Sample selection and complex effects in quantitative trait loci analysis

Purcell, Shaun

Sample Selection and Complex Effects in Quantitative Trait Loci Analysis

Shaun Purcell

A THESIS SUBMITTED TO THE UNIVERSITY OF LONDON
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Subject: Genetics

Supervisor: Prof. Pak C. Sham

Social, Genetic & Developmental Psychiatry Research Centre
King's College London

July 23, 2003

Acknowledgements

Special thanks to Pak Sham and Karestan Koenen.

Declaration

The work presented in this thesis is, to the best of my knowledge and belief, original and my own work, except as acknowledged in the text.

SHAUN PURCELL

July, 2003.

Cambridge, MA, USA.

Software

Various computer programs are referenced in this thesis: programs written by the thesis author are accessible from the following website:

`http://statgen.iop.kcl.ac.uk/qtl/`

VLADIMIR:

That passed the time.

ESTRAGON:

It would have passed in any case.

VLADIMIR:

Yes, but not so rapidly.

Beckett, 1948.

Contents

List of Tables

List of Figures

Abstract

Many statistical techniques have been developed to map genes, with great success for many aetiologically-simple traits demonstrating strong genotype–to–phenotype correspondence. More powerful and efficient approaches are still needed for complex, quantitative traits demonstrating polygenic inheritance. For such traits, that possibly involve very many genes of small individual effect, the use of selected samples is one avenue to improve power. However, although current methods can, under ideal conditions, map genes accounting for as little as 1% of variation, most complex traits will be determined by more than simple, universal and additive effects. New statistical approaches must aim to capture a range of complex effects, in which genes interact, and are correlated, with other genes and environments. Phenomena such as epistasis, gene–environment interaction and population stratification must be addressed, both to detect individual genes and, ultimately, to understand whole causal pathways. In the context of variance components quantitative trait locus linkage and association analysis, this thesis considers the optimal use of selective sampling strategies (Part I) and attempts to incorporate into QTL analysis the three complex effects mentioned above (Part II) with particular emphasis on selected samples (Part III).

Abbreviations

This list of abbreviations and conventions contains only commonly used items: items may take different meanings in different contexts (e.g. β meaning Type II error rate, or regression coefficient).

α	Type I error rate
β	Type II error rate
θ	Recombination fraction
μ	Mean, mean vector
π	Proportion of alleles shared IBD
σ^2	Variance
σ^2_A, σ^2_Q	(Additive) QTL variance
$\mathbf{\Sigma}$	Covariance matrix
χ^2	Chi-square distribution
-2LL	Minus twice log-likelihood
a	Additive genetic value
a^2	Additive genetic variance
AIC	Aikaike's Information Criterion
ASP	Affected sib pair design
B	Between sibship association
BW	Total association
c^2	Shared environmental variance
cM	Centimorgan
d	Dominance deviation
DNA	Deoxyribose nucleic acid
DZ	Dizygotic twin pair
e^2	Nonshared environmental variance

$E(X)$	Expected value of X
ED	Extreme discordant selection
EDAC	Extreme discordant and concordant selection
E-M	Expectation-Maximisation
$G \times E$	Gene-by-environment interaction
GC	Genotypic configuration
GPC	Genetic Power Calculator
H_0	Null hypothesis
H_A, H_1	Alternate hypothesis
HWE	Hardy-Weinberg equilibrium
HWD	Hardy-Weinberg disequilibrium
IBD	Identical by descent
IBS	Identical by state
$L(X)$	Likelihood of X
LCA	Latent class analysis
LD	Linkage disequilibrium
LL	Log-likelihood
LOD	Log of odds
LRT	Likelihood ratio test
LS	Least squares
MahD	Mahalanobis distance-based selection
MDis	Maximally discordant selection
ML	Maximum likelihood
MZ	Monozygotic twin pair
n^2	Nonshared residual variance
NCP	Noncentrality parameter
p, q	Allele frequency for diallelic loci
$P(X\|Y)$	Conditional probability of X given Y

PS	Proband selection
QTL	Quantitative trait locus
r_{GE}	Gene–environment correlation
s	Sibship size
s^2	Shared residual variance
SEA	SElection for Association
SEL	SElection for Linkage
SNP	Single nucleotide polymorphism
TDT	Transmission/disequilibrium test
VC	Variance components
W, W1, W2	Within sibship tests
z	Ratio of d/a
z	Probability of complete IBD sharing

Chapter 1

Introduction

1.1 The human genome and genetic variation

1.1.1 Structure of the human genome

An organism's genome is the total information carried by its genetic code, which is written in deoxyribose nucleic acid (DNA) and is necessary to build the proteins which form the basic molecular units of life. DNA is a large, complex molecule, a double-stranded nucleic acid 0.01mm in diameter – 2 metres of DNA are packed into every human cell. Each chain of nucleotide bases is made up of deoxyribose sugar attached to a phosphate group and an organic base, which will either be a purine (adenine or guanine) or pyrimidine (cytosine or thymine). The back-bone of DNA is formed by the alternating sugars and phosphates; the bases pair up with each other, adenine (A) with thymine (T), guanine (G) with cytosine (C), to make two complementary strands running in opposite directions.

In eukaryotes, DNA is organised in separate sections, called *chromosomes* - long helices of DNA surrounded by special histone proteins. Each species has a characteristic number of chromosomes: humans have 23, whilst horses have 32 and some plants have over 1,000. Male and female *gametes* (sperm and egg cells respectively)

normally contain one copy of each chromosome, and are called *haploid.* Most cells in the human body have two copies of each chromosome (i.e. one from each parent) and are called *diploid.* An exception arises with the *sex chromosomes*, X and Y: females have two X chromosomes, males have one X and one Y chromosome. The 44 other chromosomes in each human cell are called *autosomes.* The chromosomes of each pair have a characteristic length and contain similar sequences of DNA molecules; they are called *homologous* chromosomes.

Chromosomes replicate in a process called *mitosis*, during which a diploid cell divides into two diploid cells. A related process is *meiosis*, the production of the gametes, during which a diploid cell generates a haploid gamete. During meiosis, homologous chromosomes align and exchange genetic material through *recombination* to create unique haploid gametes. The probability that two positions on a chromosome are separated by a recombination event is related to the distance between them. Recombination shuffles up the genome before passing it on, and is central to the phenomena of *linkage* and *linkage disequilibrium* (or *allelic association*), which can be exploited to map genes.

Genes consist of one or more stretches of DNA that code for proteins. The particular sequence of base pairs determines the order in which *amino acids* will be joined together to form a particular protein. The basic process by which genes have their effect is described by the 'central dogma' of genetics: that DNA makes RNA makes proteins. This precludes the inheritance of acquired characteristics, which would require information to flow in the opposite direction. *Transcription* is the creation of mRNA, *translation* involves amino acids being linked up to make proteins.

Neither genome size nor number of chromosomes necessarily predict an organism's complexity or the amount of genetic information stored, as only a small proportion of the genome is functional, i.e. codes for proteins. Measured in kilobases (1 kilobase (kb) = 1000 base pairs) the human genome has 3,000,000 kb located on the

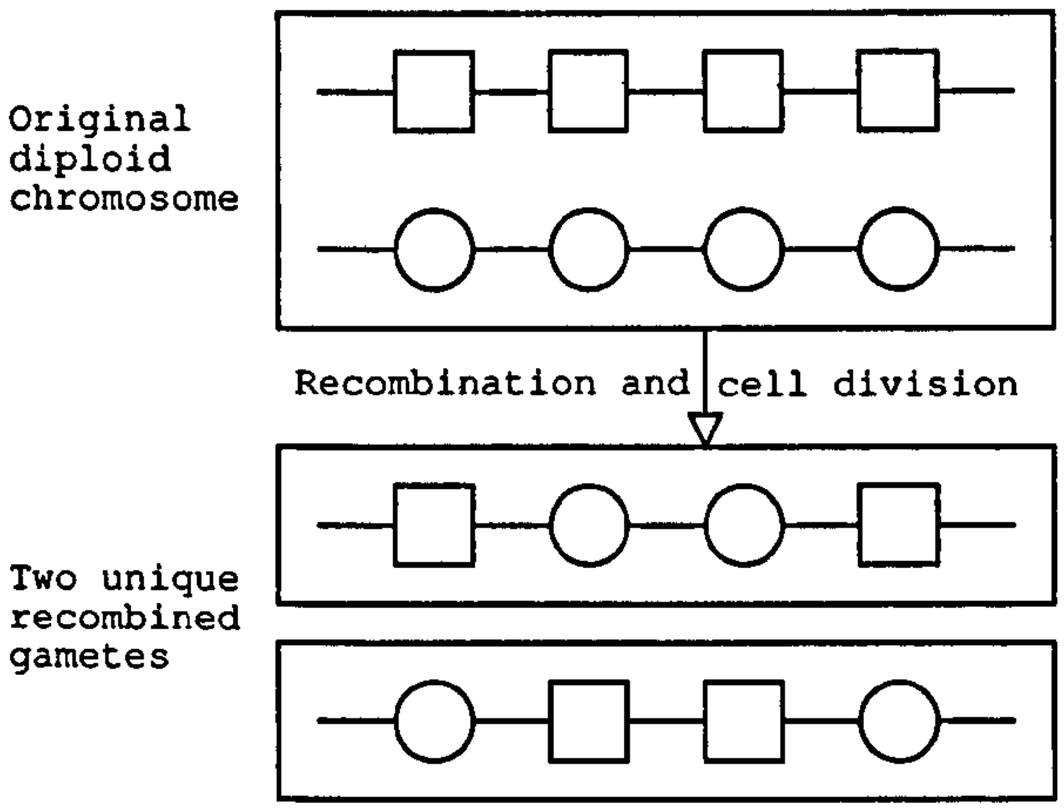


Figure 1.1: Meiosis. The top panel represents stretches of the two homologous chromosomes in a diploid cell: the squares represent regions of the paternally-inherited chromosome, the circles represent regions of the maternally-inherited chromosome. During meiosis, recombination occurs between homologous chromosomes, almost at random, such that genetic material is exchanged between homologous chromosomes to create unique chromosomes. Only one of each new homologous chromosome pair is transmitted to the gamete (see Mendel's second law of independent assortment below).

23 chromosomes, containing an estimated 30,000–40,000 genes, which may individually be anything between 1kb and 2,000kb in length. (In contrast, salamanders have 90,000,000kb on only 12 chromosomes.) Each human chromosome contains approximately 100,000kb (ranging from 250,000 on chromosome 1 to 55,000 on chromosome 21). As humans we share 98.4% of all our DNA with chimpanzees, 97.7% with gorillas.

1.1.2 Genetic variation

Although homologous chromosomes carry the same genes, they are not completely identical. Novel variation in sequence can arise from *mutation*. The most common mutation is a *base substitution*, typically of a single base (e.g. a C is replaced by a T). Mutations may also involve the insertion or deletion of genetic material, at the level of a few base pairs or even whole chromosomes.

A specific point on a chromosome is called a *locus*; variant DNA sequences at a locus are called *alleles*. If an individual has the same allele for both homologous chromosomes, the individual is *homozygous* for that allele. If an individual carries

two different alleles, the individual is said to be *heterozygous* at that locus. The combination of the two alleles is called the *genotype* at that locus.

If a variant form of a gene has a frequency in a population of at least 1%, it is typically called a *polymorphism*, as opposed to a mutation. To study individual differences is to study the polymorphic regions of the genome that are not shared by all humans. Genetic variation together with environmental variation produces all observable, *phenotypic* variation. Variation in phenotypes as diverse as blood type, height and measures of personality is, to at least some extent, caused by genetic variation.

Approximately 3 million of the 3 billion base pairs in the human genome are naturally polymorphic: most of these are *single nucleotide polymorphisms* or SNPs, which are substitutions, insertions or deletions involving only a single nucleotide. Another form of polymorphism often used in genetic studies is the short tandem repeat (STR) polymorphism: a specific sequence of bases, or motif, is repeated a variable number of times (commonly two bases, in which case the polymorphism is called a *dinucleotide repeat*).

The vast majority of this variation is unlikely to have any phenotypic effect, as it occurs within the non-coding regions of the genome. Whether or not any one mutation will have a strong, mild or nonexistent impact on a gene's product and lead to phenotypic variation depends on a number of factors (e.g. whether a substitution results in a different amino acid being produced or not – so-called *synonymous* and *non-synonymous* substitutions). Over the past two decades, increasingly detailed 'maps' of polymorphisms known to reside in particular places in the genome have been constructed. Although the majority of these *DNA markers* will not themselves have phenotypic effects, they can be used in the mapping, or *positional cloning*, of genes related to human disease.

Ultimately, knowledge of the complete genome sequence in populations of indi-

viduals will be the definitive basis for detecting previously unknown DNA variation. Presently, several other laboratory techniques can be used to screen regions for evidence of variation. Known polymorphisms can be directly detected at the DNA level. The development of the *polymerase chain reaction* (PCR) method allows exponential replication of small sequences of DNA (typically 200 – 1000 bases in length). This process results in the target DNA fragment being increased in concentration to a level at which it can be easily detected. For STR polymorphisms, this involves determining the size of the amplified fragment, which indicates the number of repeats an individual has for each homologous chromosome, i.e. their genotype. For SNPs, a technique called restriction fragment length polymorphism is commonly used, although many new technologies are emerging (e.g. micro-arrays and mini-sequencing).

The complete human genome sequence (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001) and the comparable advances in marker maps (The International SNP Map Working Group, 2001; Peltonen and McKusick, 2001) represent milestones in the accumulation of knowledge about the genome that will prove invaluable in efforts to locate disease-causing genes. How to make full use of this knowledge to direct and enhance mapping strategies is, however, a central, unresolved question in statistical genetics.

1.2 Foundations of gene-mapping

Traditional quantitative genetics (Fisher, 1918; Mather and Jinks, 1982; Falconer, 1989; Lynch and Walsh, 1998) is concerned with the aggregate properties of all genes – for example, twin studies estimate the proportion of variance attributable to all genetic effects. The complementary strategies of linkage (reviewed by Smith, 1986; Amos and de Andrade, 2001) and association (reviewed by Cardon and Bell, 2001; Jorde, 2000) allow the mapping of the chromosomal position of genes. Identification

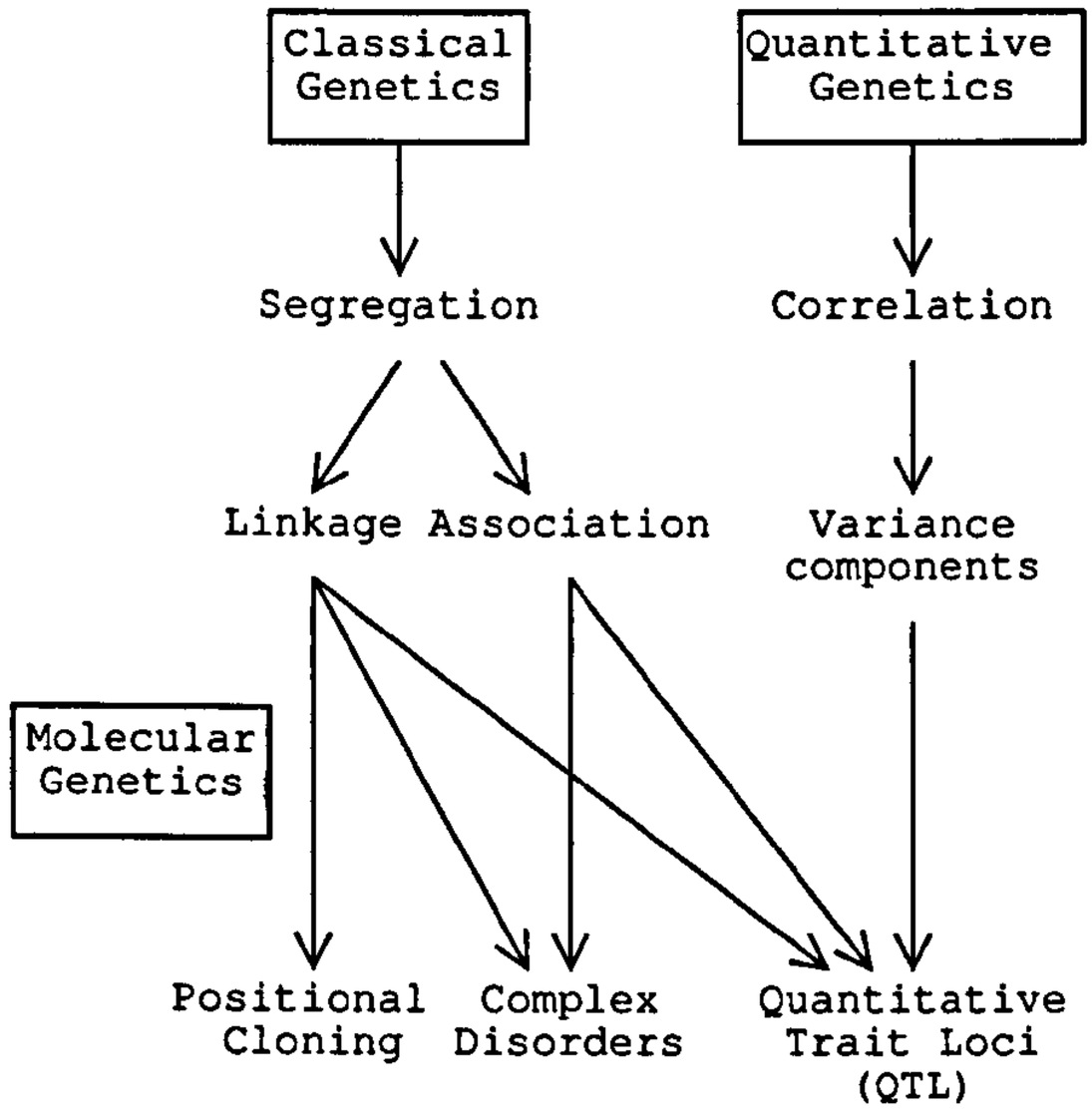


Figure 1.2: The evolution of gene–mapping.

of individual genes will offer further insights and opportunities. In the short-term, finding disease genes promises risk prediction and modification and drug-response prediction (i.e. individualised medication, based on genotype). Longer-term benefits include greater insight into the causal mechanisms and pathways of disease, and ultimately the development of new drugs and therapies.

Figure 1.2 presents an historical overview of the evolution of gene-mapping, clearly illustrating the two schools of thought that, synthesised by Sir R. A. Fisher, have lead to modern *quantitative trait loci* (QTL) genetics.

Classical genetics (exemplified by Mendel's early investigations, described below) has typically focused on discrete traits in small numbers of large pedigrees; a fundamental statistic is the *segregation ratio*. In contrast, quantitative genetics, initiated by Sir Francis Galton and Karl Pearson amongst others, studied continuous traits in large numbers of relative pairs; the fundamental statistic is the *correlation coefficient*.

Sturtevant embarked on the first gene-mapping studies in 1913, in the fruit-fly (*Drosophila melanogaster*). Gene-mapping in humans has developed only recently, largely due to the relative poverty of known genetic markers, as well as the inability

to arrange experimental human crosses. Using natural variation in DNA, however, now provides a virtually unlimited supply of genetic markers, enabling the mapping of hundreds of Mendelian (single gene) disorders, including phenylketonuria (PKU), Huntington's disease and cystic fibrosis.

This section reviews the fundamentals of the biometrical models, largely developed by Mendel and Fisher, in three inter-related sections that correspond to the three sets of parameters in classical linkage analysis: the *transmission*, *population* and *penetrance* models. The *transmission model* probabilistically describes genetic inheritance: that is, the probability of observing a genotype in offspring, given the parents' genotypes. The biological phenomenon of recombination is introduced in this context. The population model, renamed here as the *genotype model*, describes the probability distribution of genetic factors in the parents, or the *founders* as they are more generally known (i.e. individuals whose parents are not included in the study). A central concept here is *allele frequency* and the association between alleles at different loci, *linkage disequilibrium*. Finally, the penetrance model, renamed here as the *phenotype model* describes the relationship between genotype and phenotype. It is largely this third model which needs to incorporate the complex effects believed to underly most complex human traits, some of which are described in the final section of this Introduction.

1.2.1 Transmission model

Mendel's laws (Mendel, 1866) form the theoretical basis of the genetics of inheritance. Long before the discovery of DNA, Mendel concluded that a given characteristic is determined by two "factors" (i.e. genes), one of which is inherited from an individual's father, one from an individual's mother. Importantly, when any one individual passes on one of two genes to an offspring, which copy is transmitted is determined at random, and independently for different offspring. This *law of segregation* probabilistically

describes the final stage of meiosis, when the haploid gamete is formed. Mendel's second *law of independent assortment* states that the transmission probabilities are also independent between different factors. However, for genes residing on the same chromosome, the process of recombination leads to an exception to Mendel's second law – genes close together on the same chromosome will cosegregate in a dependent manner. This phenomenon forms the basis of genetic linkage.

Linkage can be observed as the tendency for certain traits to be transmitted together from generation to generation. For example, elliptocytosis and Rhesus blood group are linked. If a father has inherited both traits from one of his parents, then he will tend to transmit both traits, or neither trait, to his offspring. Thus these two traits tend to co-segregate, and this is because the traits are determined by linked genes. (Another possibility is that a single gene causes both traits, a phenomenon called *pleiotropy*).

Supposing there were no exchange of genetic material between an individual's maternal and paternal chromosomes during meiosis, then each chromosome in a gamete would be either the entire maternal or the entire paternal chromosome. In this case, all genes on one chromosome would be completely linked (i.e. co-transmitted). In this case, linkage analysis would be able to establish which genes are on different chromosomes, but it would be useless for establishing the relative position of genes on the same chromosome.

However, in reality there is an exchange of genetic material in meiosis, i.e. recombination. Each chromosome in a gamete therefore consists of alternate segments of paternal and maternal chromosome. The point where the chromosomal origin changes from maternal to paternal or vice versa is called a *cross-over*. Cross-overs occur almost at random along the genome and form the basis of a measure of genetic distance. That is, the further away two loci are, the more likely they are to be separated by a recombination event, or cross-over. Therefore, as the genetic distance between two

loci increases, linkage becomes weaker because a recombination event is more likely to have occurred. This is the basis for mapping genes by linkage.

For any two loci, the *recombination frequency* or *recombination fraction* (often labelled θ) measures this dependence in inheritance – it equals the probability that the two loci in the same haploid gamete originate from different grandparents. For loci on different chromosomes, $\theta = 0.5$. For linked loci, $\theta < 0.5$.

For two loci A and B, the two-locus genotype of an individual is A_fB_f / A_mB_m. *Haplotypes* are sets of alleles inherited from the same parent – in this case, A_fB_f is the maternal haplotype and A_mB_m is the paternal haplotype. When this individual produces gametes, one of four possible haplotypes will be transmitted; two of these will be *recombinants*:

A_fB_f	Non-recombinant
A_mB_m	Non-recombinant
A_fB_m	Recombinant
A_mB_f	Recombinant

The proportion of recombinant haplotypes is the recombination fraction (θ) between A and B, ranging from 0 (complete linkage) to 0.5 (no linkage).

Recombination fraction, genetic distance, physical distance

The expected number of cross-overs between any two loci on the same chromosome represents the *genetic map distance* (Haldance, 1919), its unit being called a Morgan (more commonly the centi-Morgan is used, the expected number of cross-overs per 100 meioses). On average, the human genome is approximately 35 Morgans in genetic distance (350 cM).

Recombination between any two loci occurs if there is an odd number of cross-overs between them during meiosis: various *map functions* have been proposed to describe the relationship between recombination fraction and 'genetic distance'. For

example, Haldane's map function is $\theta = (1 - e^{-2m})/2$ where m is the genetic distance in Morgans. Furthermore, the relationship between genetic map distance and physical distance (in DNA base pairs) is not straightforward: it varies between species, between sexes, between different chromosomes and between different regions on the same chromosome. In humans, 1 cM roughly corresponds to 1 million base pairs of DNA (1 Mb, 'megabase').

Parametric linkage analysis

The goal of standard linkage analysis is to demonstrate that a disease and a genetic marker cosegregate within families, as this will imply the presence of a disease-causing gene in the broad region of the chromosome containing the marker. Although it is possible to narrow this region by typing more markers on more individuals, the resolution of linkage analysis is fundamentally limited by the number of recombination events that can be observed. For example, for two loci separated by a recombination fraction of 0.01, it is necessary to observe 100 meiosis to have above 50% chance of observing just 1 recombination.

Under certain favourable conditions it is possible to determine with certainty whether or not a recombination event has occurred: in this case, the recombination fraction is simply calculated as the proportion of recombinant gametes observed. In the most simple case, one would consider the offspring of double heterozygote A_1B_1/A_2B_2 and double homozygote $A_1A_1B_2B_2$ parents. The *phase* of the double heterozygote genotype is known (indicated by the "/"), indicating that A_1 and B_1 originated from the same ancestral chromosome, as did A_2 and B_2, i.e. as opposed to the A_1B_2 and A_2B_1 haplotypes. In this case, the offspring haplotypes A_1B_1 and A_2B_2 are non-recombinant, whereas the offspring haplotypes A_1B_2 and A_2B_1 are recombinant. In this example, loci A and B are both genetic loci – more typically, one of the loci would in fact be a *disease locus* representing the affection status of the

individual. Given a simple mode of genetic inheritance, (e.g. dominant, recessive – described below under the phenotype model section) the recombination fraction between the disease locus and the marker locus could be directly estimated by counting the number of recombinant and non-recombinant gametes.

More sophisticated statistical methods are usually applied to pedigree data in order to estimate the recombination fraction between a marker and putative disease locus. The method of *maximum likelihood* (Fisher, 1922) is most often used: probability models are formulated in terms of various unknown parameters (e.g. the recombination fraction in this case). The *likelihood* can be calculated, which is the probability of the observed data given the parameter value(s). The difference in likelihoods for different parameter values provides a measure of relative support for those different parameter values, conditional on the observed data. The set of parameter values which gives the highest likelihood are the *maximum likelihood estimates* (MLE) of the parameters. The likelihood of the data under the estimated value of θ, $\hat{\theta}$, is compared to the likelihood of the data under the hypothesis of no linkage, i.e. $\theta = 0.5$. The common log of the likelihood ratio gives the *lod score* (Morton, 1955):

$$LOD(\theta = \hat{\theta}) = \log_{10} \frac{L(X|\theta = \hat{\theta})}{L(X|\theta = 0.5)}$$

The interpretation of a lod score of 3 at $\theta = 0.1$, for example, is that the data are 1000 times more likely to have arisen if $\theta = 0.1$ as opposed to $\theta = 0.5$. Lod scores (for particular values of θ) can be summed across families and studies to produce summary test statistics. Morton (1955) suggested that a cumulative lod score of 3 represents strong evidence for linkage; a cumulative lod score of -2 represents strong evidence against linkage.

In many other applications, log-likelihoods are calculated using natural logarithms (i.e. base e), in which case twice the difference in log-likelihood will asymptotically

follow a χ^2 distribution. A lod score can be multiplied by $2 \ln 10$ to give a χ^2 statistic: a lod score of 3 corresponds to a χ^2 of 13.8 which has a significance value of $p = 0.0001$ (one-tailed test). Allowing for the multiple testing inherent in genome-wide scans, and the low prior probability that any one test locus is linked to a QTL, this corresponds to a *genome-wide significance level* of approximately 0.05. For Mendelian disorders, this criterion appears to be valid – Rao et al. (1979) found that 98% of linkages with a lod score of 3 or more were replicated in subsequent studies. For complex traits, more stringent significance criteria are called for (Lander and Kruglyak, 1995).

Nonparametric linkage analysis

A drawback with parametric linkage analysis is that many potentially unknown factors must be specified for the trait model. For Mendelian disorders, for which *segregation analysis* can estimate the mode of transmission from data on affection status in pedigrees, this may well be practicable. However, for complex traits that may show multiple genetic effects, interactions, heterogeneity, and so on, this position becomes increasingly undesirable. *Nonparametric* linkage analyses (often called *allele sharing* methods), first introduced by Penrose (1935), differ from parametric linkage analyses in that an explicit phenotype model need not be specified (see below). In nonparametric methods, marker allele sharing between relatives is correlated with trait-similarity between relatives. 'Allele sharing' can be defined in two ways: where alleles are *identical by state* (IBS) or *identical by descent* (IBD). Two alleles are IBS simply if they contain the same DNA sequence; to be also IBD, the two alleles must have descended from a single allele in a recent common ancestor. The definition of allele sharing based on IBD and not IBS has proven more powerful and robust, and forms the basis of the nonparametric approach to linkage used in this thesis.

Assessing IBD therefore describes the patterns of co-inheritance of chromosomal regions between related individuals. At any one locus, a sibling pair can share 0,

Sib 1	Sib 2	IBD
A_1A_3	A_1A_3	2
A_1A_3	A_1A_4	1
A_1A_3	A_2A_3	1
A_1A_3	A_2A_4	0
A_1A_4	A_1A_3	1
A_1A_4	A_1A_4	2
A_1A_4	A_2A_3	0
A_1A_4	A_2A_4	1
A_2A_3	A_1A_3	1
A_2A_3	A_1A_4	0
A_2A_3	A_2A_3	2
A_2A_3	A_2A_4	1
A_2A_4	A_1A_3	0
A_2A_4	A_1A_4	1
A_2A_4	A_2A_3	1
A_2A_4	A_2A_4	2

Table 1.1: The 16 identity-by-descent (IBD) configurations for a sibling pair.

1 or 2 alleles IBD. The *IBD distribution* refers to the probabilities of these IBD values. If paternal and maternal genotypes are A_1A_2 and A_3A_4 respectively, there are four equally likely offspring genotypes: A_1A_3, A_1A_4, A_2A_3 and A_2A_4. These four genotypes are transmitted to each sibling independently, giving sixteen combinations (or *inheritance vectors*) all of equal probability 1/16, as shown in Table 1.1.

Inspection of Table 1.1 shows IBD 0, 1 and 2 occurring 4, 8 and 4 times respectively, giving the 'prior probabilities' of IBD sharing of 1/4, 1/2 and 1/4. That is, for any given locus, 25% of randomly selected sib pairs will have inherited the same alleles from both their father and mother. The mean proportion of alleles shared IBD for full siblings (often called π) is therefore $\frac{0}{2} \times 0.25 + \frac{1}{2} \times 0.50 + \frac{2}{2} \times 0.25 = 0.5$.

In the Table 1.1 example, it would be possible to calculate IBD at the test marker locus given the sibling and parental genotypes. This is often not the case: for example, if all four parental alleles are not unambiguously identifiable (e.g. a parental

mating type of $A_1A_2 \times A_1A_1$) or were completely missing (e.g. only the siblings were genotyped) then statistical methods are needed to infer IBD sharing at that locus. Typically, multiple markers are used to more accurately infer IBD sharing at all points along each chromosome, in *multipoint* or *interval mapping* approaches. Two types of recursive procedure are commonly used: the Elston–Stewart algorithm (Elston and Stewart, 1971) that can handle large pedigrees but only small numbers of loci and the Lander–Green algorithm (Lander and Green, 1987) that can handle a large number of loci but only small pedigrees.

The simplest nonparametric linkage test for a binary disease trait is the affected sibling pair (ASP) method (Suarez et al., 1978). If IBD can be unambiguously inferred from the marker genotype data, then the test of linkage is simply whether or not the average proportion of alleles shared IBD at the test locus is greater than 50% (the expected value for full sibling pairs, under the null of no linkage). If IBD information is incomplete, it can be estimated using a likelihood-based 'maximum lod score' (MLS) method (Risch, 1990)

Classical linkage methods were developed around Mendelian disorders, and so are primarily aimed at mapping binary disease traits, i.e. those measured on a 'yes'/'no' scale. Many complex traits are better defined in terms of a quantitative phenotype rather than a binary category, however. Some disorders may represent the high end of a continuum, with no well-defined threshold, in which case directly measuring the continuum may provide more power. Alternatively, some phenotypes, such as height or IQ, are truly continuous in nature. A locus that contributes to variation in a continuous trait is called a *quantitative trait locus* (QTL). There are two main classes of nonparametric QTL linkage test in common use, both of which are described below in more detail: Haseman-Elston regression (Haseman and Elston, 1972) and variance–components models (e.g. Amos, 1994; Fulker et al., 1999).

1.2.2 Genotype model

Mendel's laws and biology provide a strong model of genetic transmission, i.e. of *intra-familial* factors. That is, conditional on the founder member genotypes, the probability distribution of offspring genotypes is known. It is often necessary to construct probability models for *inter-familial* differences also: roughly speaking, the probability of the founders possessing certain genotypes.

Consider, for simplicity, a *diallelic* locus (i.e. a polymorphism with only two alleles, A and a). The *allele frequency* of one allele, say A, is often labelled p while the frequency of the a allele is $q = 1 - p$. Under random mating in large populations, the *genotype frequencies* are p^2, $2pq$ and q^2 for genotypes AA, Aa and aa respectively. This frequency distribution of genotypes is called *Hardy-Weinberg equilibrium.* For multiple loci, the *haplotype frequency* will be the product of the constituent allele frequencies if the alleles are independent in the population. If the haplotype frequency differs from the product of the allele frequencies, then the alleles are said to be in linkage disequilibrium (described below).

An allele is associated with a disease if the allele frequency is higher in affected than unaffected individuals. Alternatively, associations may be based on specific genotypes or haplotypes rather than specific alleles. For quantitative traits, association results in extreme-scoring individuals being more likely to possess particular alleles / genotypes / haplotypes.

Basic association study design

Samples of unrelated individuals are often collected for association studies. For a binary trait, the most simple association study design is the case-control design, as commonly used in epidemiology. The 'risk factor' might be either an allele, a genotype or a haplotype – a frequency difference between cases and controls is evidence for association. A contingency table (e.g. a 2×2 table of allele by disease status with

$2N$ observations – N individuals, each with two alleles) could be constructed, and the association tested with a χ^2 test of independence. Alternatively, logistic regression could be applied, with caseness as the dependent variable. When considering multiple tightly linked loci, some form of *haplotype analysis* is usually adopted, which looks for associations between specific haplotypes and disease. This can be more powerful than single allelic associations in some circumstances, although often it is not possible to unambiguously determine an individual's haplotype, and so haplotype frequencies have to be estimated.

To improve efficiency, *DNA pooling* is proving an increasingly popular design: the DNA from all cases forms a single pool, the DNA from all controls forms a second pool. The allele frequencies in the two pools can be measured by molecular techniques (Daniels et al., 1998) although only two PCRs are required (instead of one per individual). The main drawback of this design is that only the main effects of individual alleles can be considered: it is not possible to consider differences in genotype or haplotype frequencies between cases and controls. Nevertheless, as a screening instrument, DNA pooling offers a great deal of promise.

If the outcome variable is a quantitative trait, simple regression-based methods can be applied to test for association: for example, with the trait as the dependent variable and genotype as the independent variable (coded 1, 0 and -1, for example, for the three genotypes). Dominance effects and multiple (> 2) alleles can be included by adding further dummy variables, as can covariates. Alternative methods include likelihood based approaches (see Chapter 9).

Family-based association designs

Population stratification (described below) is a potentially severe problem for any association study. Two types of solution have been proposed to counter this: using family-based association methods and, more recently, using information from individ-

uals' genetic backgrounds.

The most popular family-based association tests are based on parent-offspring trios, although many extensions have been subsequently proposed. The haplotype-based haplotype relative risk (HHRR) test (Terwilliger and Ott, 1992) and the *transmission / disequilibrium test* (TDT) (Spielman et al., 1993) both analyse trios according to which alleles heterozygous parents transmit to affected offspring. For example, if parental genotypes are Mm and mm and the offspring genotype is Mm, then the 'control genotype' is mm (i.e. constructed from the untransmitted parental alleles). Although these tests are robust to stratification, they require three genotypes for every case-control pair; also, parents might not be available, especially for late-onset disorders.

The original family-based tests focus on binary disease traits, although quantitative versions have since been developed. The TDT was in fact originally introduced as a test of linkage; however, as the TDT depends also on the presence of association, its properties are similar to other tests of association (with the exception of robustness to stratification) and it is now generally regarded as a test of association.

An alternative family-based association test was proposed by Fulker et al. (1999). The test is designed for sibships (with or without parental genotypes) and is primarily designed for quantitative traits. The model is framed in a variance-components framework and incorporates a simultaneous test for linkage and as well as association. This model forms the basis for much of this thesis, and is described below in more detail.

Linkage disequilibrium mapping

The tests of association described above can be used to determine whether the test locus is a causal variant or not. A study design that seeks to establish this is called a *candidate gene* design.

Ignoring stratification, significant results might not be due to the alleles at the test locus having an effect, however: a second nearby locus could be the functional variant. This phenomenon is sometimes called indirect association, due to *linkage disequilibrium* and can be utilised in genome-wide association studies. That is, anonymous marker loci across the genome can be used to test for both linkage and association: the phenomena of linkage and linkage disequilibrium will mean that any marker linked to the disease locus (very tightly linked in the case of association) will potentially retain some of the signal from the disease locus.

Linkage disequilibrium (LD) refers to the association of two alleles at different loci. When a new mutation first arises and is passed down the generations, the chromosomal background upon which that mutation occurs will also tend to be co-inherited along with the mutant allele. This will lead to a correlation between the new mutant allele and whatever alleles happen to be on the same chromosome when the mutation occurs. Recombination during meiosis is constantly reshuffling the genome, and so tends to break down LD between between loci. However, as very tightly linked loci are unlikely to be separared by recombination, the rate of decay of LD will be slower the closer together two loci are. Because of this, LD can be used to localise genes.

Let M and D represent specific alleles at a marker and disease locus: if the probability of observing the haplotype MD equals the product of the probability of observing M and D individually, then M and D are in *linkage equilibrium* (i.e. statistically independent). Deviation from this measures LD, often represented as $\delta = P(MD) - P(M)P(D)$. LD reduces after n generations by a factor of $(1-\theta)^n$, the probability of M and D not being separated by a recombination event over all n generations. Therefore, for very tightly linked loci, LD can persist for very many generations. The theoretical predictions have been largely supported by empirical studies (e.g. Abecasis et al., 2001b), although this area is a complicated and fast-

expanding field of enquiry.

Both linkage and association analysis therefore rely on the biological phenomenon of recombination to provide a measure of genetic distance: in linkage analysis, recombination events are inferred for the meioses in the pedigrees under study; in association analysis, the cumulative effect of unobserved recombination over many generations generates the profile of linkage disequilibrium.

1.2.3 Phenotype model

The final component of a genetic model describes the relationship between genotype and phenotype. The *penetrance* or *phenotype* model essentially characterises the observable effects of genetic variation. As mentioned, the majority of phenotypes are measured either as binary diseases or continuous dimensions, although a character might be measured in both ways (e.g. a clinical diagnosis of depression versus a severity index of depressive symptoms). For binary traits, Mendel outlined the modes of major gene action; the work on continuous traits was inspired by Sir Francis Galton, a relative of Darwin. A synthesis of the two traditions was eventually reached by Fisher, who outlined a simple biometrical model to explain continuous variation in terms of particulate inheritance.

For a particular trait, a *dominant* allele is one that expresses itself at the expense of an alternate allele. Conversely, a *recessive* allele is one whose expression is suppressed by a dominant allele. For a binary trait and two-allele system D/d where D is the disease-causing allele, if D is dominant then a single copy is sufficient to cause disease; if D is recessive then two copies are necessary to cause disease. Table 1.2 illustrates the case of dominance. In this case, the genotypic ratio in offspring of heterozygous parents is $1 : 2 : 1$ for genotypes $DD : Dd : dd$; the phenotypic ratio is $3 : 1$ for disease:no disease. This example assumes *fully penetrant* genotypes. Penetrance is the probability of developing disease conditional on genotype. In the example given

	Paternal	
Maternal	D	d
D	DD (Disease)	Dd (Disease)
d	Dd (Disease)	dd (No disease)

Table 1.2: An example of dominant transmission: a single copy of the dominant D allele is sufficient to cause disease.

in Table 1.2, $P(\text{disease}|\text{DD}) = \text{P}(\text{disease}|\text{Dd}) = 1$ whilst $P(\text{disease}|\text{dd}) = 0$.

In Mendelian disorders there is a direct relationship between genotype and pathology. A *complex* trait is one that exhibits familial clustering (suggesting at least some genetic component) but does not, if it is a binary trait, occur in Mendelian proportions in pedigrees. Departures from the basic Mendelian model include *reduced penetrance* (the absence of disease in individuals with the disease genotype) and *phenocopies* (the presence of disease in individuals without the disease genotype). Another form of complex disease is *locus heterogeneity*: where a disease is linked to the test marker only in a proportion of families (a different gene causes disease in the unlinked families). A second recessive disease locus segregating only in a minority of families might explain phenocopies for a major dominant locus.

Reduced penetrance and phenocopies can be seen as 'exceptions to the rule' and incorporated within the standard linkage framework by allowing each genotype to have a specific penetrance value between 0 and 1 – this is often called the *single major locus* (SML) model. In order to perform parametric linkage analysis, it is generally necessary to specify these penetrances (as well as the allele frequencies of the unobserved disease locus).

Most genuinely complex traits are expected to depart from the single major locus model to a much greater extent. Truly complex traits are highly *multifactorial* – "of many factors". These factors include genes, environments and interactions; interactions may be between genes, between environments, or between genes and environments. Mendel's laws of inheritance still apply to the many individual genes that influence complex traits. If the only factor to influence of trait were a single

dialleic QTL, for which there are only 3 possible genotypes, there would only be 3 possible phenotypes. If two QTL were operating, there would be 9 possible two-locus genotypes; with 5 QTL there are 243 possible genotype combinations. In this way, and combined with environmental variation, a relatively small number of factors can result in near-continuous variation.

For complex traits, the genotype–phenotype relationship is a probabilistic or statistical one. Furthermore, the genetic architecture of the trait cannot be adequately described in terms of a single major locus. For binary diseases, risk now stems from normal variation of normal genes and not from abnormal mutation. For continuous traits, extreme-scoring individuals are more likely to possess certain genes over other genes.

It is common to describe the genetic architecture of complex traits in terms of the cumulative effects of the multiple unobserved loci. The most important summary statistic is *heritability*, the proportion of trait variation attributable to genetic variation. As shall be seen, genetic variation can be indirectly estimated without measuring any specific loci, by comparing phenotypic similarity in groups of related individuals that differ in genetic similarity.

Mendelian disorders are typically rare: whilst the impact on the individual may be great (many Mendelian disorders are fatal), the burden of disease in public health terms is small. In contrast, complex diseases tend to show the opposite pattern, being common with massive public health implications. Examples of complex diseases and traits include coronary heart disease, hypertension, diabetes, obesity, anxiety and depression. Whilst parametric linkage analysis has had great success with rare, single-gene Mendelian disorders, results for complex traits have been a trail of modest lod scores over very broad regions that fail to replicate.

1.3 Modern QTL mapping methods

Broadly speaking, the development of modern methods to detect and locate QTL for complex traits proceeds from the intersection of population genetics, experimental design and statistical analysis.

Population genetics: In general, the transmission model, as outlined above, is biologically well-understood and statistically well-characterised; the same cannot be said for the genotype model. To this end, population genetics addresses a whole range of complex issues concerning the genotypic landscape within which humans exist. Central questions include the 'allelic spectrum' associated with disease: whether the *common disease / common variant* (CD/CV) hypothesis will hold, as opposed to most disease being caused by a massively heterogeneous mixture of very rare mutations (Weiss and Terwilliger, 2000). The answer to this question will have potentially massive implications for the success of current mapping strategies and future directions. The structure of linkage disequilibrium in the human genome is beginning to be empirically addressed (e.g. Daly et al., 2001; Abecasis et al., 2001b), as are the related questions of population size, structure and history (e.g. Rosenberg et al., 2003).

Experimental design: There are many experimental design issues in QTL mapping. A central question has addressed the relative merits of linkage versus association mapping, which has been largely resolved (Risch and Merikangas, 1996; Sham et al., 2000b). Other issues include optimal pedigree size and structure, marker type and density and sample selection strategies. Questions concerning the definition of the phenotype may be more difficult to resolve, although the use of repeated and/or multiple measures is a promising start (Boomsma, 1996).

Statistical analysis: A great deal of recent work has explored statistical and analytical issues in QTL mapping. The relative merits of parametric and nonparametric approaches are still debated. Bayesian estimation methods are being developed and implemented in accessible packages (e.g. WinBUGS, Gilks et al., 1994) that comple-

ment the traditional use of maximum likelihood estimation. One important analytic question concerns the development of tests that are robust in non-normal data and selected samples.

These three areas are represented in this thesis in different contexts: for example, Chapters 6 and 9, which address the detection and correction of population stratification effects in tests of association, are largely based on previous population genetic work. Chapters 2 and 3 consider the use of selected samples, an issue of experimental design. Chapters 3, 9 and others, use a novel 'conditioning-on-trait-value' analytic approach.

The remainder of this section provides an overview of variance components methods in quantitative genetics, including twin, linkage and association analysis.

1.3.1 Variance components methods

Variance components (VC) methods decompose phenotypic variance into a number of components depending on the type of data available. Developed by Fisher in 1918, the basic model illustrates how the variance of a continuous trait can be decomposed into additive and non-additive factors, in a manner compatible with both Darwinian evolutionary theory and Mendelian genetic theory. Subsequently, variance components models in statistical genetics have grown to include multiple genetic and environmental factors, as well as a number of complex interacting and covarying phenomena, e.g. covariates, multiple traits, different pedigree types, assortative mating, sibling interaction, gene-by-environment interaction (Hopper, 1993).

The sources of variance in VC methods may represent measured (e.g. *DRD*4 genotype) or unmeasured (e.g. additive polygenic effects) variables. Typically, 'environmental' factors are unmeasured, defined to represent everything 'non-genetic', including age, sex, and measurement error as well as more traditional environmental factors, e.g. household or cultural factors.

Dealing with unmeasured, or *latent* factors, Fisher showed that $\sigma^2_G = \sigma^2_A + \sigma^2_D$, where σ^2_G represents the genetic variance (due to either a single locus or many loci) which is decomposed into orthogonal additive and dominance components, σ^2_A and σ^2_D. By combining the joint genotypic distributions of different relative types with simple expressions for the mean and variance of a hypothetical QTL effect, Fisher derived expressions for additive and dominance QTL variance components in terms of allele frequencies and genotypic effects, and the correlations of these components between relatives. Using the modern notation of Falconer (1989), the three genotypic means of a diallelic QTL are often expressed as $m+a$, $m+d$ and $m-a$ respectively, which gives the general formulae:

$$\sigma^2_A = 2pq\,[a + d(p-q)]^2$$

and

$$\sigma^2_D = (2pqd)^2$$

where the *additive genetic value*, a, represents the additive effect of the locus, and is twice the difference between the two homozygotes; the *dominance deviation*, d, represents the dominant, or non-additive, effects of the locus and is the difference between the heterozygote and the midpoint of the two homozygotes.

Fitting VC models

VC methods developed for decomposing correlational data into genetic and environmental components have been extended for QTL linkage and association analysis (Schork, 1993; Amos, 1994; Kruglyak and Lander, 1995a; Fulker and Cherny, 1996; Almasy and Blangero, 1998; Fulker et al., 1999).

In basic twin analysis, phenotypic variance is typically decomposed into a part attributable to *additive genetic effects* across all unobserved polygenes, a part at-

tributable to *common environmental effects* (i.e. shared between twins) and a part attributable to *nonshared environmental effects* (i.e. not shared between twins). Such a model is typically labelled the ACE model (A, C and E corresponding to the three genetic and environmental components just mentioned).

In QTL linkage analysis, phenotypic variance is decomposed into a part attributable to linkage to individual marker loci ("QTL variance"), and residual parts due to polygenes and environmental effects. That is, rather than modelling the specific allele frequencies and penetrances of the trait locus, as in parametric linkage analysis, only the resultant variance it causes is considered.

In QTL association analysis, the QTL is modelled as a fixed effect in the means model. Variance components representing residual sources of variation, and sibling covariation, are often added to the basic model.

For both QTL linkage and association, often only additive QTL effects are considered, so tests involve only a single parameter (although dominance effects can easily be incorporated). The QTL variance component or fixed effect is estimated at the candidate locus, or in the case of a genome scan, at each point along the genome using interval mapping approaches.

Model-fitting attempts to match observed data (either in the form of summary statistics such as means, variances and covariances, or as raw data) with their expected values, which are derived from theoretical models containing these sources of variation. Maximum likelihood is the criterion most commonly used to fit model expectations to data. Fixing a model parameter to zero is equivalent to dropping that particular term from the model. By comparing nested models, *likelihood ratio tests* (LRT) of parameters can be constructed (i.e. comparing a model in which a parameter is freely estimated against one in which it is fixed to be 0, for instance). The likelihood ratio test provides a test of significance for the dropped components.

Although VC models may contain any number of components, not all compo-

nents will necessarily be identifiable in a given data set. For example, from MZ and DZ reared-together twin data alone, it is not possible to estimate variance due to additive *and* dominance genetic effects, shared and nonshared environmental effects simultaneously. In this case, the model is not *identified.*

General VC methodologies that use maximum likelihood estimation and allow for different pedigree structures and data missing at random were first introduced in statistical genetics in the 1970s (Lange et al., 1976; Thompson, 1977a,b). Subsequently, computer packages have been developed for easy and flexible implementation of such models, including the *Mx* package (Neale, 1997).

In general, the log-likelihood of the vector of observed trait values $\mathbf{x_i} = [x_{i1}, x_{i2}, ..., x_{is}]$, for the i^{th} pedigree containing s members, is (ignoring the constant)

$$\ln L_i = -\frac{1}{2}\left[\ln|\mathbf{\Sigma_i}| - (\mathbf{x_i} - \mu_i)'\mathbf{\Sigma_i^{-1}}(\mathbf{x_i} - \mu_i)\right]$$

where the trait has a multivariate normal distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. Both μ and $\mathbf{\Sigma}$ can be defined by different parameters (corresponding to fixed and random effects respectively). Maximum likelihood models typically assume multivariate normality – deviations from this assumption can lead to problems in the power and robustness of the test, as reviewed below.

Twin data can be modelled with $\mu = \begin{bmatrix} m & m \end{bmatrix}$ and covariance structure

$$[\mathbf{\Sigma_{MZ}}]_{ij} = \begin{cases} \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2 & \text{for i = j} \\ \sigma_A^2 + \sigma_D^2 + \sigma_C^2 & \text{for i} \neq \text{j} \end{cases}$$

for MZ twins and

$$[\mathbf{\Sigma_{DZ}}]_{ij} = \begin{cases} \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2 & \text{for i = j} \\ \frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + \sigma_C^2 & \text{for i} \neq \text{j} \end{cases}$$

for DZ twins (where i and j index each twin). The polygenic additive and dominance variance components are σ^2_A and σ^2_D; the shared and nonshared environmental components are σ^2_C and σ^2_E. With only MZ and DZ data, this model is not identified – it is usual practice to constrain either σ^2_D or σ^2_C to zero.

1.3.2 QTL linkage analysis

Just as twins can be classified according to zygosity, a population of sibling pairs can be classified into 3 groups on the basis of IBD sharing (0, 1 or 2 alleles) at the position of a putative QTL. The phenotypic variance due to the additive and dominance effects *at the QTL* (i.e. instead of polygenic effects) will be shared as follows for the 3 groups:

	Components of variance shared between full siblings	
IBD	Additive	Dominance
0	0	0
1	$\frac{1}{2}\sigma^2_A$	0
2	σ^2_A	σ^2_D

As a result, IBD status at the QTL correlates with phenotypic sibling similarity. If a marker locus is in linkage with the QTL, then IBD status at the marker will be positively correlated with IBD status at the QTL. If the recombination fraction between marker and QTL is θ, then the correlation in IBD status is $(1-2\theta)^2$. Therefore, when $\theta = 0.5$ (marker and QTL unlinked) the correlation is 0; when $\theta = 0$ (marker actually is the QTL) the correlation is 1. Under linkage, therefore, increased allele sharing between siblings at the marker is related to increased allele sharing at the QTL, which in turn is related to increased phenotypic similarity. The test of linkage is therefore whether or not there is any correlation between allele sharing at the marker and phenotypic similarity. This is typically implemented by dropping the QTL variance terms from the model (i.e. fixing σ^2_A and σ^2_D to 0). This approach can be extended to deal with larger sibships and general pedigrees.

A pair's IBD status can be measured in two ways: 1) by three values representing the probability of sharing either 0, 1 or 2 alleles IBD, z_0, z_1 and z_2, or alternatively, 2) by the proportion of alleles shared IBD, i.e. $\pi = z_1/2 + z_2$ representing additive effects and z_2 representing dominance effects. In the first case, often called the "weighted-likelihood" approach, the likelihood is a mixture of three models, $L_W = z_0 L_0 + z_1 L_1 + z_2 L_2$ where L_0, for example, represents the model when IBD is 0. In the second case, often called the "pi-hat" approach, the likelihood is simply $L_{\hat{\pi}}$ where the covariance term for the single model is $\hat{\pi}\sigma_Q^2 + \sigma_S^2$. Under simple conditions, the two models are equivalent; the weighted-likelihood approach presents computational problems when larger pedigrees are considered (i.e. it is necessary to consider every possible IBD sharing configuration for the whole pedigree). However, the pi-hat approach can lead to problems when missing IBD information is imputed (Dolan et al., 1999).

1.3.3 QTL association analysis

Linkage and association are complementary methods (Elston, 1998; Suarez and Hampe, 1994; Monks et al., 1998). In general, linkage is able to detect only major effects, but over large distances, whereas association is able to detect minor effects but only over small regions. Linkage always leads to association, although for most loci this association is purely intra-familial, i.e. there is no association at the population level (Hodge and Elston, 1994). On the other hand, association may or may not be due to linkage. A systematic linkage genome scan may be conducted with only several hundred markers; to cover the entire genome using association-based methods may take thousands or even tens of thousands of markers, however. There is a growing interest in systematic genome-wide association analysis – driving this trend are recent developments in DNA pooling and multi-locus haplotype analysis, and the use of unlinked background markers to protect against spurious association.

In VC models, association is modelled in the means vector as a fixed effect of

genotype. For a diallelic locus, for sibling i, the additive effects A_i are coded 1, 0 and -1 for test locus genotypes GG, Gg and gg respectively; dominance effects D_i are coded 0, 1 and 0. The means vector for full sibling pairs is therefore

$$\mu = \begin{bmatrix} aA_1 + dD_1 & aA_2 + dD_2 \end{bmatrix}$$

and the test for association is between a model in which $a = d = 0$ as opposed to a model in which they are freely estimated (d can be fixed to 0 in both models to provide a 1 degree of freedom test of additive effects only). The effect of IBD sharing at the test locus can still be modelled in the covariance structure: in this way it is possible to construct models that determine whether an association explains all of the linkage, i.e. to ask whether or not the test locus is the causal variant or merely in linkage disequilibrium with it (by testing for linkage whilst simultaneously modelling the association). Extensions to this model that make a test robust to population stratification effects have been developed (Fulker et al., 1999) and extended to general pedigrees (Abecasis et al., 2000). This 'between–within' model features extensively in Chapter 3 and is described there in more detail. Essentially, the within component (looking at intrafamilial association) is robust to population stratification effects whereas the between component (looking at inter-familial association) is not.

1.4 Statistical power

1.4.1 Hypothesis testing and error rates

Most research in the behavioural sciences is dominated by Fisherian hypothesis testing, in which the usual aim of research is to reject a *null-hypothesis*, e.g. that no difference exists between two group means. Rejecting the null-hypothesis represents evidence in favour of the research hypothesis, e.g. that a difference exists. This stan-

dard of proof is probabilistic: typically, a threshold which would only be expected to be passed by chance 5% of the time *if the null-hypothesis were true* is taken as the criterion for proof. Such a *significance criterion* is an arbitrary convention; often referred to as α, it represents the probability of a false rejection of the null.

The Fisherian method provides a deterministic and objective way of making a 'yes or no' decision given a set of appropriate data, as to whether a theory is supported or not. The output of a test is a p value, which is the probability of a result at least as large as the one observed occurring by chance if the null were true. As mentioned, a 5% level of chance is usually taken to be a reasonable criterion, which of course implies that 5% of results will be spurious. The p value is not the probability of the null hypothesis being true, as it is commonly misunderstood. Also, as Fisher emphasised, the nonrejection of the null does not assert its truth: nonsignificant results do not necessarily support the conclusion that 'no difference exists'.

Whereas Fisher's system only specifies one null hypothesis, that is either rejected or not, Neyman & Pearson (1928a, 1928b) developed a system that chooses between two hypotheses. The *alternate hypothesis* specifies a precise, non-null state of affairs and has an associated risk of error, β. The two types of inferential error can be clearly identified within their formulation: false-positive and false-negative errors. A false-positive, or *Type I error*, represents the rejection of a true null (occurring at rate α); a false-negative, or *Type II error*, represents the failure to reject the null when the alternate hypothesis is true (occurring at rate β). The *power* of a statistical test is a measure of its ability to find a difference when one exists, the probability of rejecting a false null-hypothesis, or $1 - \beta$.

Power, significance criterion, sample size and magnitude of effect are functionally interdependent variables. Power can be described, in terms of the other three variables, as "the probability of detecting a given effect size in a population, from a sample of size N, using a significance criterion α". In this way, for an hypothesised

effect size, power can be determined prior to conducting an experiment, by choosing appropriate values for α and N. Alternatively, post hoc power calculations may be useful aids when interpreting results, particularly when nonsignificant results have been obtained. For example, one could ask what the minimal detectable effect size was for a given power, α and N. Also, a nonrejection of the null is only meaningful if power is high, i.e. any effect present would have been likely to be detected. If power is low (so there was never a reasonable chance of rejecting the null) then a negative result should not be regarded as definitively discrediting the research hypothesis. Only if power is set to 95% (i.e. $\beta = 0.05$) then the nonrejection of the null can be, with reasonable confidence, taken as evidence for the truth of the null.

By convention, many researchers accept a power of 80% as a sensible goal: less power would result in an unreasonably high Type II error rate, whereas greater power would often entail an unreasonably large sample size. Of course, the adequacy of the power level will also depend on the effect size; other issues such as multiple testing and the prior probability of an effect being present will also determine choice of power. These other considerations play a particularly strong role in many statistical genetic applications.

The power of a test can be increased by raising either significance criterion or sample size. Good experimental design, which increases the effective effect size is a further way of increasing power. Through increased measurement accuracy, effect size is a quantity at least partially under the experimenter's control: anything that reduces error variance (more reliable measures, better designs, more appropriate statistics, especially multivariate, incorporation of co-variates) will increase power.

The consequences of chronic low statistical power being the norm in a research domain are sobering. If power is on average only marginally greater than α, then a large number of published studies may well be Type I errors. Average power around the 50% level yields a pattern of inconsistent replication. Unfortunately, a great deal

of time and money has been spent on poorly designed experiments that, at best, stand little chance of doing what they are supposed to, and, at worst, are advancing Type I errors in the literature.

1.4.2 Power of QTL linkage and association analysis

In general, linkage designs are excellent for high penetrance genes – classical parametric linkage analysis of Mendelian traits can proceed with just a few large multiplex families. Power diminishes gradually with genetic distance: for example, there is only a small reduction at 5cM, so a large portion of the genome can be covered with a relatively small number of markers. Linkage is unaffected by allelic heterogeneity between families; however, power is poor for low penetrance genes.

For QTL linkage, the information content of a sibling (or other relative) pair is determined by the overall sibling (relative) correlation, the proportion of variance the QTL accounts for, the variance in IBD sharing (determined by the informativeness of the marker loci) and the trait values of the pair (Sham et al., 2000b). The information content of a general pedigree depends on the number and type of related individuals present (Rijsdijk et al., 2001): it is approximately equal to the sum of all pairwise combinations, so that large sibships and large pedigrees are generally informative and efficient. Assuming complete marker information and that the test locus is the QTL, the expected *noncentrality parameter* (NCP) of the variance components linkage test for a sibship of size s is approximately (Sham et al., 2000b)

$$\lambda_L \approx \frac{s(s-1)}{2}\left(\frac{1}{8}V_A^2 + \frac{3}{16}V_D^2 + \frac{1}{4}V_A V_D\right).$$

where V_A and V_D represent the proportions of variance accounted for by the additive and dominance QTL effects (exact solutions are also given). The NCP plus the degrees of freedom equals the χ^2 likelihood ratio test statistic. Unlike power, the NCP is linear

with sample size, and so is a natural unit with which to compare the properties of different methods. Essentially, the NCP represents an amalgam of effect size and sample size. The above equation shows that the NCP for the test of QTL linkage is proportional to the *square* of QTL heritability, indicating the low power to detect genes of small effect.

For the variance components association test, the NCP for the between and within sibship association tests are (Sham et al., 2000b)

$$\lambda_B \approx \frac{\frac{s+1}{2}V_A + \frac{s+3}{4}V_D}{sV_S + V_N}$$

and

$$\lambda_W \approx (s-1)\frac{\frac{1}{2}V_A + \frac{3}{4}V_D}{V_N}$$

The NCP of the between sibship test is up to 3 times greater than the within NCP, although the between test is not robust to population stratification. The presence of residual shared variance increases the NCP of the within test. In contrast to linkage, the NCP is proportional to QTL heritability, indicating the greater power to detect genes of small effect.

These power calculations have been extended to include the effects of incomplete linkage or linkage disequilibrium between the marker and the QTL (Sham et al., 2000b). Along with other tests (e.g. TDT and case-control tests for discrete traits), these calculations can be performed automatically using the Genetic Power Calculator web tool (Purcell et al., 2003)[1].

1.4.3 Calculating power

Consider the following example, for a simple case-control study design. The data are the frequency of a risk factor in 30 cases and 30 controls; the test of independence

[1]Located at `http://statgen.iop.kcl.ac.uk/gpc/`

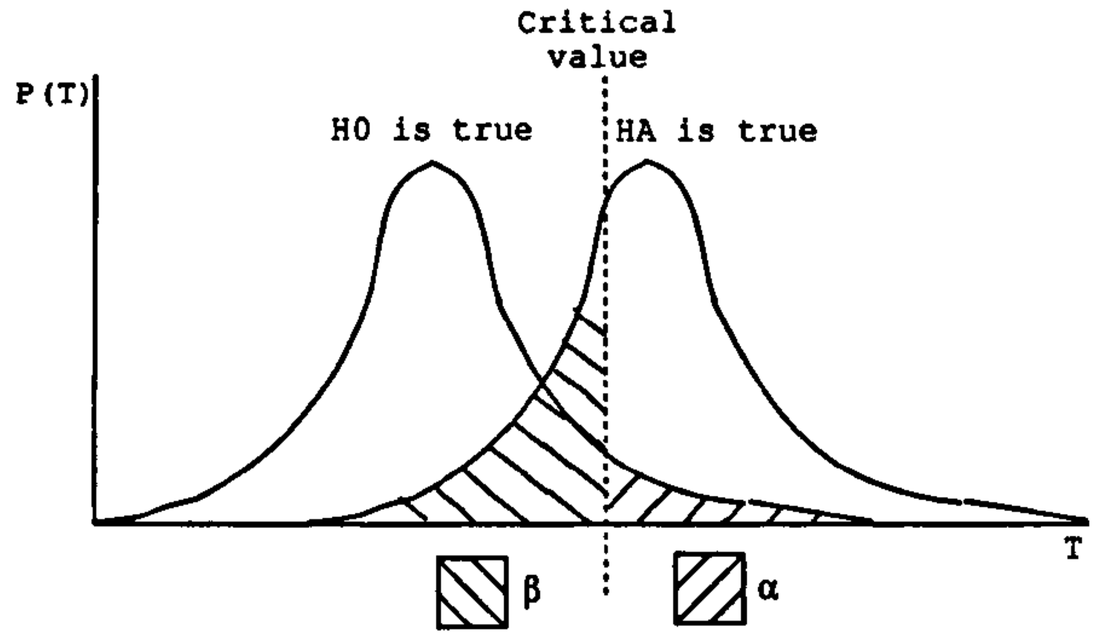


Figure 1.3: Type I and Type II error rates: the two curves represent the distributions of the test statistic under the null and alternate hypotheses. The shared areas under the curve give α and β, which are also functions of the critical value.

follows a χ^2_1 distribution. For a given sample size, hypothetical effect and type I error rate, we wish to calculate power. The first step is to calculate the expected χ^2 value, which determines the distribution under the alternate hypothesis (see Figure 1.3). Imagine the effect is such that the expected sample frequencies are:

	Case	Control
Risk allele present	20	10
Risk allele absent	10	20

The χ^2 test statistic is $\sum \frac{(O-E)^2}{E}$ where O are the observed cell frequencies and E are the expected cell frequencies assuming independence, which gives a value of 6.666. The second step is to calculate the critical value for the desired type I error rate, α. Setting $\alpha = 0.05$, the inverse central χ^2 distribution function gives the critical value: this gives the critical value X given the area under the curve for a central χ^2 (i.e. NCP=0). For $\alpha = 0.05$, for a 1 df χ^2, the critical value is 3.84146. Finally, the noncentral χ^2 distribution function gives the power: this gives the area under the alternate curve given the NCP (the expected χ^2 of the test) and above the critical value X, which equals 0.73. Therefore, the power to detect an association given the effect and sample size of the above data is 73%.

1.5 The analysis of selected samples

As mentioned above, maximum likelihood estimation is almost always based on relatively strong assumptions of multivariate normality. If the trait is non-normal in the population, this can influence the test statistics: in particular, tests on means are influenced by skewness, whereas tests on variances are influenced by kurtosis. As Allison et al. (1999b) point out, as well as intrinsically non-normal traits, factors such as the presence of a major gene (not linked to the markers under study), some types of gene–environment interaction and the use of binary phenotypes all induce non-normality. Furthermore, even if a trait is normally-distributed in the population, it will not be in a selected sample.

In a comprehensive assessment of the robustness of the variance components approach to QTL linkage, Allison et al. (1999b) illustrate that certain forms of non-normality and selective sampling can indeed inflate false-positive rates, and discuss some of the potential solutions. Various approaches to the analysis of data which violate the distributional assumptions of maximum-likelihood analysis have been proposed: adjusting the test statistic by a 'deflation factor' (Blangero et al., 2000); modelling using a non-normal distribution (e.g. t distribution Lange et al., 1989); use of generalised estimating equations, quasi-likelihood methods, M-estimation and other robust methods (e.g. Huggins, 1993); nonparametric approaches (Kruglyak and Lander, 1995b); Monte-Carlo Markov Chain methods (e.g. Guerra et al., 1999); finally, transformation and data-trimming approaches (e.g. Wang et al., 1998). Even transforming the data prior to analysis is not always guaranteed to achieve normality of the error distribution; furthermore, such a procedure could potentially reduce power (e.g. in the presence of a pseudo-major gene effect that effectively generates a mixture distribution of scores).

Most of these alternative methods are designed to analyse mildly non-normal data, or samples with extreme outliers: they do not necessarily address the potentially more

		Full sample	Select on X	Select on Y
H_0	β	0.000	0.001	0.000
	SE	0.022	0.025	0.063
	p	0.501	0.503	0.500
	Type I ($\alpha = 0.01$)	0.008	0.011	0.009
H_A	β	0.499	0.499	0.778
	SE	0.019	0.020	0.032
	p	0.000	0.000	0.000

Table 1.3: Impact of sample selection on regression estimates: results for the full sample, or for samples selected on either the independent variable (X) or the dependent variable (Y).

severe problem of selected samples. Sample selection can cause problems even for otherwise relatively robust methods such as ordinary least squares regression. Table 1.3 illustrates the impact of selection on a basic (non-genetic) regression analysis of Y on X. A sample of 5000 observations was simulated 10,000 times. Both variables are simulated to be normally-distributed: under the alternate hypothesis H_A, the regression coefficient β should be 0.5. Selection on X retains only observations where X is more than 1 standard deviation away from the mean, likewise for Y. Under the null H_0, correct Type I error rates are obtained. The standard error is higher when selecting on Y however, although the average p value and type I error rate are correct in this simple scenario. Under the alternate H_A, the regression coefficient β is biased when selecting on Y. In cases more complex than this simple univariate regression, or when using alternate analytic methods such as maximum likelihood estimation, it is possible to obtain inflated Type I error rates under the null when selecting on Y (e.g. see Chapter 3).

It is known that standard VC approaches can produce inflated test statistics and increase false-positive results when applied to samples selected for phenotypic extremes (Allison et al., 1999b; Dolan et al., 1999). If the phenotypic scores for individuals not genotyped are available, it is possible to incorporate them into the analysis and impute prior IBD probabilities, although this will only work if a weighted-likelihood (as opposed to pi-hat) approach is adopted (Dolan et al., 1999). Problems with this

approach are that the phenotype data must be available, it might be hard to generalise to larger pedigrees (being based on the weighted likelihood approach), and it does not address the additional issue of non-normal population distributions.

Other solutions to the selected sample problem include the use of ordinary least squares regression procedures (Haseman and Elston, 1972) and permutation tests (e.g. Dunn et al., 1993). The Haseman-Elston (H-E) linkage method relies on regressing the squared sibling-pair trait difference on the proportion of alleles shared IBD at the marker locus. A negative slope suggests linkage because it correlates similarity at the trait locus with similarity at the marker locus. For sibling pairs (Sham and Purcell, 2001) and general pedigrees (Sham et al., 2002b) an extended form of the H-E method has been shown to have similar power to VC linkage analysis. One interesting advantage is that this H-E approach might be more robust than standard VC linkage analysis. Chapter 8 illustrates a two-locus extension of the extended H-E method in selected samples.

Permutation tests involve deriving critical values for the likelihood ratio test statistic via random resampling techniques, instead of assuming that it follows a particular distributional form (e.g. a 50:50 mixture of χ^2_1 and 0). Such a procedure will always ensure correct type I error rates, although it may not be optimally powerful.

1.5.1 Conditioning on trait values

A further possibility is to adjust the likelihood by the probability of the ascertainment event (e.g. Elston and Sobel, 1979, based on the *ascertainment correction* used in segregation analysis (Morton and MacLean, 1974)). This involves dividing the standard likelihood by the probability that the proband(s) falls into a specified ascertainment region R, e.g. having a score above a certain threshold

$$L_A(X|G) = \frac{\sum_\pi L(X|\pi)P(\pi|G)}{\int_R L(X)}$$

For large families, or cases where the ascertainment scheme is more complex than simple threshold-based proband selection (i.e. as outlined in Chapter 2) this approach can be difficult. An alternative is to condition on the actual observed trait values, which does not require any knowledge of the ascertainment scheme. That is, $\int_R L(X)$ is replaced with $L(X)$ (e.g. Hopper and Matthews, 1982; Ewens and Shute, 1986). This 'conditioning-on-trait-values' approach is equivalent to a new class of method that models genotype conditional on phenotype, rather than phenotype conditional on genotype as is commonly done (e.g. Alcais and Abel, 1999; Dudoit and Speed, 1999, 2000; Sham et al., 2000a). It may be more robust to model genotype conditional on trait in the presence of both non-normality and selected samples. This is because samples tend to be selected on the basis of phenotype rather than genotype – selecting on a dependent variable can cause problems as seen in the simple regression example above, whereas selecting on the independent variable is generally valid. These methods follow the general approach advocated by Risch and Zhang (1995) in the analysis of pairs selected for extreme discordance.

Whereas Dudoit and Speed (2000) use a score statistic to test for linkage, Sham et al. (2000a) have implemented a conditional test within the more standard maximum likelihood variance components framework: this latter approach is adopted in this thesis, although conceptually the two approaches should be identical. In the context of both linkage and association tests, the general approach outlined in Sham et al. (2000a) is used in Chapters 2, 3, 7 and 9 of this thesis.

As stated above, the covariance matrix for full sibling pairs is

$$[\Sigma_{\mathbf{FS}}]_{ij} = \begin{cases} \sigma_Q^2 + \sigma_S^2 + \sigma_N^2 & \text{for i} = \text{j} \\ \pi\sigma_Q^2 + \sigma_S^2 & \text{for i} \neq \text{j} \end{cases}$$

where σ_Q^2 is the QTL variance, and σ_S^2 and σ_N^2 are the shared and nonshared residual variance components. This covariance matrix can be re-expressed in terms of a single

parameter σ^2_Q, if values for the variance v and sibling correlation r are fixed prior to analysis (i.e. based on the population values, which must be either estimated from the full sample or based on prior knowledge).

$$[\Sigma_{FS}]_{ij} = \begin{cases} v & \text{for i} = \text{j} \\ rv + (\pi - 0.5)\sigma^2_Q & \text{for i} \neq \text{j} \end{cases}$$

Using the weighted-likelihood approach, the standard likelihood of the QTL linkage test can be written

$$L(X|G) = \sum_{\pi} L(X|\pi)P(\pi|G)$$

where $P(\pi|G)$ represents the IBD probability for 0, 1 or 2 allele sharing, estimated conditional on the marker data G; the normal density function gives $L(X|\pi)$.

The principle of the conditioning-on-trait-values approach is to re-express the likelihood as the probability of the marker data conditional on trait. Using Bayes Theorem,

$$L(G|X) = \frac{L(X|G)P(G)}{L(X)} \propto \frac{\sum_{\pi} L(X|\pi)P(\pi|G)}{\sum_{\pi} L(X|\pi)P(\pi)}$$

For sibling-pair linkage, this method maintains the correct type I error rate for phenotypically selected and non-normal samples. Full power is retained for selected samples from a normal distribution; power for non-normal samples can be slightly attenuated (Sham et al., 2000a). It is possible that such a procedure will be more powerful than a permutation test when the correct model is known, although effects of mis-specifying the correct model need to be addressed by simulation studies, as in Chapter 3. Specifying the model involves fixing the mean, variance and covariance to their population values (which is also required in the new H-E approaches mentioned above).

Figure 1.4 plots $L(G|X)$ and $L(X|G)$ for unrelated individuals. The left column represents a diallelic QTL with additive genetic value $a = 1$, dominance deviation

$d = 0$ and allele frequency $p = 0.5$. The right column represents $a = 1$, $d = 0.5$ and $p = 0.1$. The top row of graphs represent $L(X|G)$; the middle row represent $L(X) = \sum_G L(X|G)L(G)$; the bottom row represent the conditional likelihood $L(G|X) = L(X|G)/L(X)$. Under different conditions the conditional likelihood is expected to have different properties in comparison to the unconditional likelihood: these issues are explored further in specific Chapters of the thesis.

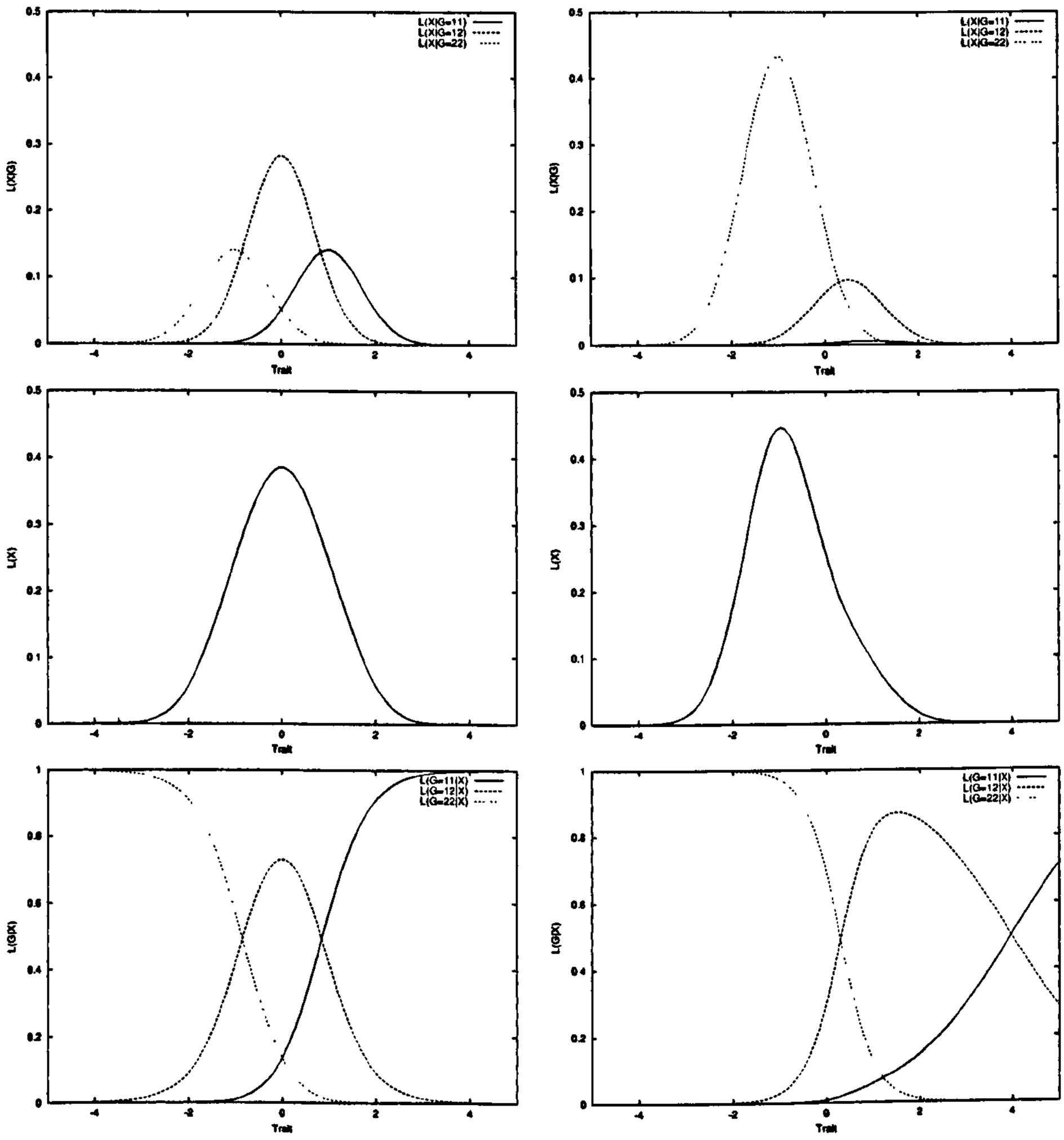


Figure 1.4: Unconditional and conditional likelihoods: left column represents an additive QTL effect, right column illustrates a QTL with dominance. Rows (top to bottom) show $L(X|G)$, $L(X)$ and $L(G|X)$ respectively.

1.6 Complex traits and effects

The concept of *genetic architecture* is a property of a specific phenotype in a specific population, rather than a biological universal. In this sense, it is a 'moving target' that will vary according to gene frequencies, environmental factors and other factors including age and sex. Critical genetic factors may differ between groups, either quantitatively or qualitatively. Such conditions would inevitably lead to the patterns of non-replication and weak signals found in current complex trait genetic studies, although low statistical power is an alternative explanation.

As many review articles in the last decade have commented, the techniques described so far will undoubtedbly face many challenges from the inconvenient realities of the genetics of complex human traits (e.g. Lander and Schork, 1994; Plomin et al., 1994; Risch and Merikangas, 1996). In this final section of the Introduction, three types of complex effect are considered: gene–environment interaction, epistasis and population stratification.

1.6.1 Gene × environment interaction

Simple quantitative genetic models average over any group differences within a population. The presence of gene–environment interaction ($G \times E$) will mean that a single statistic is no longer adequate to describe a whole population, as genetic effects will now depend on individuals' environments. A heritability of 50%, for example, might equally entail scenario S_1 where, for all individuals, differences in the trait are equally due to genetic and environmental factors or scenario S_2 where, for half the population, the trait is completely genetically determined, whereas for the other half the trait is completely environmentally determined. In the context of twin analysis, consideration of $G \times E$ aims to distinguish between scenarios such as S_1 and S_2. This requires that the E component of the $G \times E$ is a measured variable that indexes the

differential aetiologies present in S_2. For example, if the 50:50 split reflected males and females, this would represent a $G \times$ sex interaction.

It is possible to detect $G \times E$ within various study designs (Heath et al., 2002); G and E can be either latent or measured variables. When both G and E are latent variables, it is possible to detect $G \times E$ as a heteroscedastic bivariate twin distribution, where twin pair difference correlates with twin pair sum (Jinks and Fulker, 1970). However, as well as suffering from low power, this test also is sensitive to non-normality in the trait. More importantly, beyond indicating that *some* form of interaction is occurring, it sheds no light on underlying processes. Having both G and E as measured variables provides the most power for detecting $G \times E$; the results will potentially be very informative also, beginning to map onto the underlying biology. For example, sex moderates the effect of the *APOE* $e4$ allele on cognitive decline, where women show higher $e4$-associated risk than men (Yaffe et al., 2000). Additionally, the $e4$ allele moderates the impact of estrogen in women on cognitive decline, as the estrogen use is associated with less cognitive decline only in women without the $e4$ risk allele. Chapter 7 outlines a model for this kind of $G \times E$.

Chapter 4 considers the case of latent $G \times$ measured E, which is most relevant to the classical twin study. For example, additive genetic effects on depression symptoms interact with marital status in women, where unmarried women show greater levels of genetic influence (Heath et al., 1998). Another example of latent $G \times$ measured E is that a religious upbringing seems to attenuate genetic influences on the personality trait of disinhibition (Boomsma et al., 1999). Testing for $G \times E$ with a binary moderator such as marital status is straightforward. The parameters of interest (e.g. a^2, c^2 and e^2) are estimated for "exposed" and "unexposed" individuals separately. A test of $G \times E$ is given by equating the parameters across exposure group and observing the associated decline in model fit (i.e. testing for heterogeneity).

Continuous moderator variables

Complex human traits are often best defined in quantitative terms, to avoid the potential loss in power associated with artificial dichotomisation of a continuous variable. Many potential moderator variables are also most naturally defined in quantitative terms: some obvious examples include age, gestational age, socio-economic status, educational level, consumption of food, drugs or alcohol. Although typical approaches to $G \times E$ are often limited to binary moderators, it is equally possible to allow for continuous moderating variables that may differ between twins in a pair.

The most basic $G \times E$ interaction involving a continuous moderating E variable implies that genetic effects increase or decrease as a linear function of the moderator. Although this formulation covers a large class of $G \times E$, a second nonlinear class is also considered in Chapter 4, where genetic effects may, for example, be attenuated at extreme high *and* extreme low levels of a moderator.

Gene–environment correlation

$G \times E$ is often conceptualised as genetic control of sensitivity to different environments. A related phenomenon, G–E correlation (r_{GE}) represents genetic control of exposure to different environments (Kendler and Eaves, 1986). Equivalently, of course, $G \times E$ is the environmental control of differential gene effects, whereas r_{GE} is the environmental control of gene frequency. A recent example of r_{GE} showed that genetic influences on alcohol and drug misuse are correlated with various aspects of the family and school environment (Jang et al., 2001) and we might expect r_{GE} to feature in many other complex traits. Typical approaches to $G \times E$ in twin analyses involving stratification of a sample by the environmental moderator variable (Neale and Cardon, 1992) have been unable to disentangle $G \times E$ and r_{GE} in a single analysis, however. For example, if individuals in a certain environment show greater genetic influence, this could be due to either (1) the environment modifying the effects of certain genes or (2) certain

trait-influencing genes being more likely to be present in that environment. A method described in Chapter 4 is able to discriminate between these alternatives and to allow analysis of $G \times E$ in the presence of r_{GE}.

Unmodelled $G \times E$ and r_{GE}

If not explicitly modelled, $G \times E$ and r_{GE} will impact on standard twin models, in terms of biased parameter estimates. In short, interaction between A and C acts like A; interaction between A and E acts like E. Correlation between A and C acts like C; correlation between A and E acts like A. For example, in the case of $A \times C$ interaction, if a standardised trait $T = aA + cC + iAC + eE$ then the expected variance is $\mathrm{Var}(T) = a^2 + c^2 + i^2 + e^2$, assuming that the latent variables A, C and E have unit variance. The expected twin covariances are

$$\begin{aligned}
\mathrm{Cov}(T_1, T_2) &= a^2\mathrm{Cov}(A_1, A_2) + c^2\mathrm{Cov}(C_1, C_2) + e^2\mathrm{Cov}(E_1, E_2) + i^2\mathrm{Cov}(A_1C_1, A_2C_2) \\
&= a^2 + c^2 + i^2 \quad \text{for MZ twins} \\
&= a^2/2 + c^2 + i^2/2 \quad \text{for DZ twins}
\end{aligned}$$

as $\mathrm{Cov}(A_1, A_2)$ is 1 for MZ twins, 0.5 for DZ twins; $\mathrm{Cov}(C_1, C_2) = 1$ and $\mathrm{Cov}(E_1, E_2) = 0$ for all twins; also $\mathrm{Cov}(A_1C_1, A_2C_2) = \mathrm{Cov}(A_1, A_2)\mathrm{Cov}(C_1, C_2) = \mathrm{Cov}(A_1, A_2)$. Similar covariance algebra can show that $A \times E$ interaction contributes to the E component.

If A is correlated with (rather than interacting with) an environmental variable, say C, with correlation r_{AC} then the expected trait variance is $Var(T) = a^2 + c^2 +$

$2ac \times r_{AC} + e^2$ and the expected twin covariances are

$$\begin{aligned}\text{Cov}(T_1, T_2) &= a^2\text{Cov}(A_1, A_2) + c^2\text{Cov}(C_1, C_2) + e^2\text{Cov}(E_1, E_2) + ac\text{Cov}(A_1, C_2) \\ &\quad + ac\text{Cov}(A_2, C_1) \\ &= a^2 + c^2 + 2ac \times r_{AC} \quad \text{for MZ twins} \\ &= a^2/2 + c^2 + 2ac \times r_{AC} \quad \text{for DZ twins}\end{aligned}$$

as $\text{Cov}(A_1, C_2) = \text{Cov}(A_2, C_1) = r_{AC}$. Similarly, if A and E are non-independent then

$$\begin{aligned}\text{Cov}(T_1, T_2) &= a^2 + c^2 + 2ae \times r_{AE} \quad \text{for MZ twins} \\ &= a^2/2 + c^2 + ae \times r_{AE} \quad \text{for DZ twins}\end{aligned}$$

1.6.2 Epistasis

Epistatic interaction represents the modification of allelic and genotypic effects at one locus contingent upon the genotype at a different locus. Equivalently, epistasis occurs when the combined effect of two or more genes on a phenotype could not have been predicted as a sum of their separate effects.

For two loci, A and B, analysed separately, it is possible that alleles at A may show associations with a trait while alleles at B do not. Joint analysis, however, may reveal evidence for epistasis between the two loci: it is possible, for example, that alleles at locus B might modify the effect of the alleles at locus A. Therefore, the possibility of seemingly unrelated loci actually playing crucial roles in the aetiology of complex traits is a consequence of epistasis.

In general, *additive* genetic effects occur when alleles at a locus and across loci simply and independently sum to result in a net phenotypic effect. In contrast, effects of an allele which are modified by the presence of other alleles (either at the same locus or at different loci) are *nonadditive genetic* effects. In particular, an allele ×

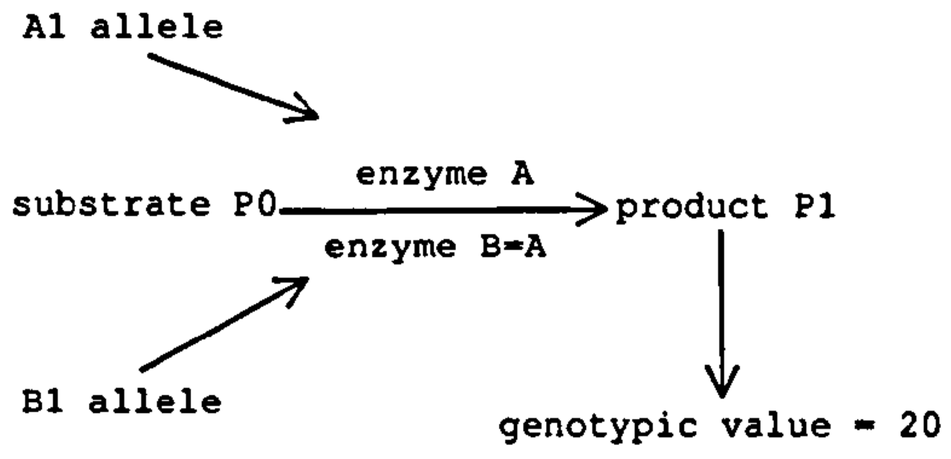


Figure 1.5: Example of duplicate gene action at the level of biochemical pathway

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	20	20	20
B_1B_2	20	20	20
B_2B_2	20	20	0

Table 1.4: Duplicate gene action at two loci

allele interaction occurring between two alleles at the same locus is called *dominance* whereas an interaction occurring between the alleles (two or more) at different loci is called *epistasis*. Dominance is therefore sometimes called *intra-locus* interaction whereas epistasis involves *inter-locus* interaction.

Inter-locus nonaddivity, or epistasis, might, for example, result from interaction at a biochemical level between two gene products. Figure 1.5 illustrates so-called *duplicate* gene interaction. In this example, there are two duplicated loci essentially serving the same function: producing an identical enzyme required to produce product P_1 and resulting observable phenotype. Thus, if either locus A or locus B produces a functional gene product, the enzymatic pathway functions correctly. In this way, the effect of one gene can effectively mask the effect of the other: it is only when both genes are homozygous for a recessive, non-functioning allele that the biochemical process is not completed and a different phenotype is produced. Assuming that the alternate alleles at the two loci, A_2 and B_2, do not result in gene product P_1, then we might expect to observe the following genotypic values as shown in Table 1.4. This would often be called dominant–dominant duplicate gene action.

Alternatively, two genes may code for enzymes that function at different points in the same pathway, such that both gene products are needed to produce the final

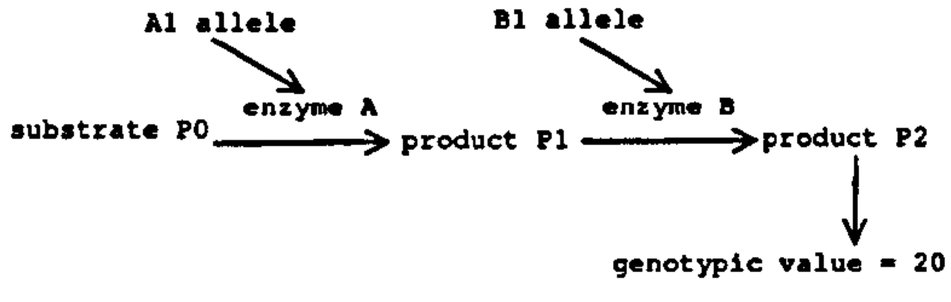


Figure 1.6: Example of complementary gene action at two loci

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	20	20	0
B_1B_2	20	20	0
B_2B_2	0	0	0

Table 1.5: Complementary gene action at two loci

product. This is called *complementary* gene interaction (Figure 1.6): if either gene is non-functioning, then the final product of the pathway is not produced. The corresponding matrix of genotypic means is represented in Table 1.5. As the A_1 and B_1 alleles are acting in a dominant manner, this table describes the scenario where an individual needs at least one A_1 allele and at least one B_1 allele to have the 'normal' phenotypic value of 20 (assuming that a phenotypic value of 0 corresponds to 'disease', for example).

These two models equally represent recessive gene interaction, e.g. if the phenotypic labels were reversed, such that 0 was 'normal' and 20 was 'diseased'. Dominant × dominant duplicate gene action is equivalent to recessive × recessive complementary gene interaction if the phenotypic 'direction' is reversed; dominant × dominant complementary epistasis is equivalent to recessive × recessive duplicate epistasis.

Figure 1.7 shows a more complex example of a system that would produce strong epistatic interaction. In this case, the genotypic values will depend on *pairs* of alleles across the two loci occurring together: A_1/B_1 and A_2/B_2 are functioning allele pairs whereas A_1/B_2 and A_2/B_1 are non-functioning pairs. In the absence of dominance, this pattern of results is described as additive × additive epistasis. No single allele is any longer associated with higher phenotypic values: there are no increaser or decreaser alleles, only increaser and decreaser *combinations*. The corresponding table

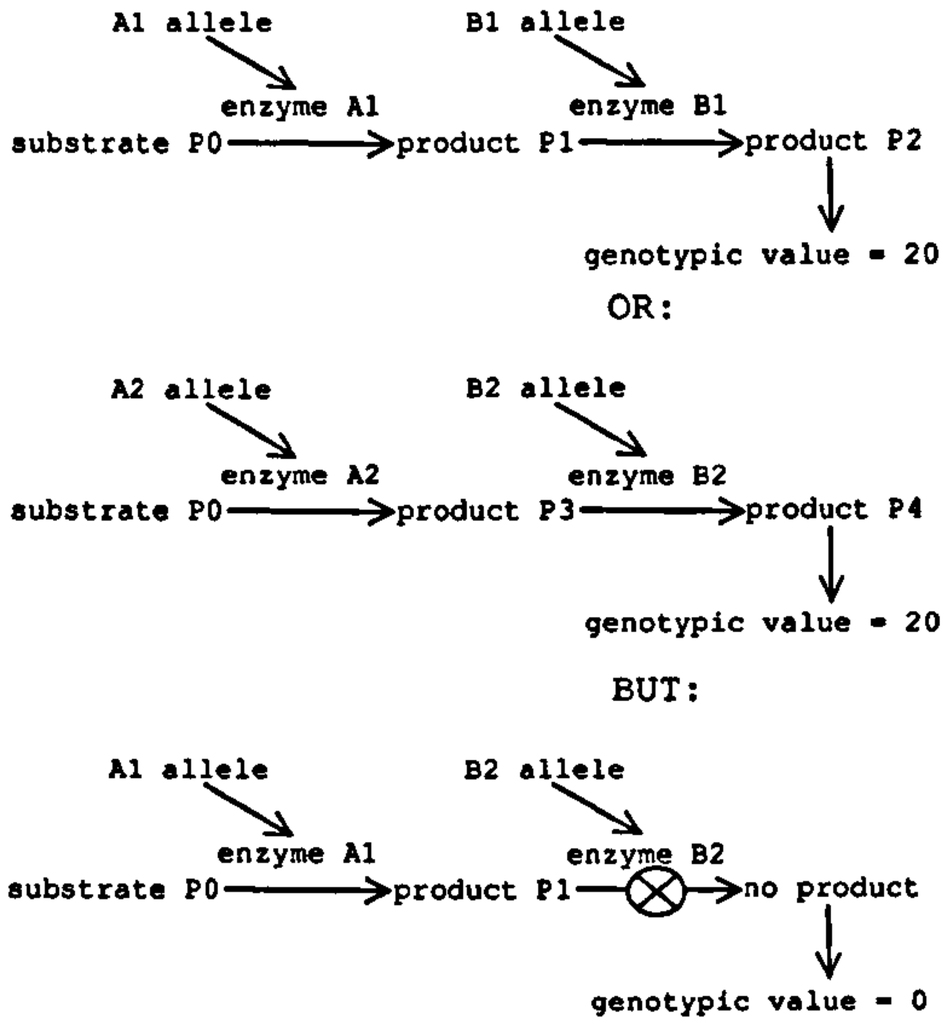


Figure 1.7: Example of complex gene interaction at the level of biochemical pathway

	A_1A_1	A_1A_2	A_2A_2
B_1B_1	20	10	0
B_1B_2	10	10	10
B_2B_2	0	10	20

Table 1.6: More complex gene interaction at two loci

of genotypic means might be as in Table 1.6. Considering individuals with the A_1A_1 genotype, the effect of the B_1 allele is to increase the phenotypic value. In contrast, B_1 is associated with *lower* phenotypic values in individuals with the A_2A_2 genotype, whereas for individuals heterozygous at locus A, there is no effect of the B locus at all.

In the context of parametric linkage, a number of studies have suggested that single-locus models allowing for reduced penetrance often perform as well as two-locus models allowing for epistasis (e.g. Vieland et al., 1993). Two-locus linkage models (e.g. Schork, 1993) are problematic in many ways: for example, many unknown parameters must be specified (standard parametric two-locus linkage analysis with two diallelic loci requires 14 parameters) whilst examining every pairwise combination of markers will result in multiple testing issues. Nonetheless, two-locus models have been successfully applied. For example, for multiple sclerosis (MS), Tienari et al.

(1994) found evidence for epistasis between the MBP gene on chromosome 18 and the HLA complex on chromosome 6. Chapters 5 and 8 consider models of epistasis for quantitative traits.

1.6.3 Population stratification

Population stratification refers to a recent admixture of subpopulations which may differ in allele frequencies at many loci across the genome. A stratified sample is therefore one in which discrete subpopulations that do not interbreed as a single randomly-mating unit are pooled together. Differences in allele frequencies between subpopulations may then give rise to several effects, including a deviation from Hardy-Weinberg equilibrium (HWE) sometimes known as the "Wahlund effect", a decrease in observed heterozygosity. The early population genetic work on population stratification was primarily concerned with its impact on population structure and the evolutionary process (Wright, 1951), although Li (1969) highlighted its potential impact in disease-gene association studies. If cases and controls are not matched for ethnic background, population stratification effects can lead to spurious association. Although the primary focus was on population stratification generating type I, or false positive errors, stratification can also reduce power (that is, to increase type II errors) if the stratification effect 'masks' the trait locus effect.

In practice, there have been few clear examples of when population stratification has actually lead to a "spurious association" (Thomas and Witte, 2001). One often-cited example is of non-insulin dependent diabetes in the Pima and Papago Native American tribes and a haplotype at the immunoglobulin G locus, where an effect of proportion of recent European ancestry was observed (Knowler et al., 1988). However, in general there are often great difficulties replicating associations (Terwilliger and Weiss, 1998) and it is unclear to what extent stratification may play a role here, given that we have not been able to accurately measure stratification in a sample until

recently.

In the 1990s many researchers were beginning to turn from linkage to association based strategies to detect genes of small effect for complex traits (Risch and Merikangas, 1996). To address concerns over possible hidden stratification effects, a series of family-based tests of association were developed, including the transmission disequilibrium test (TDT) (Spielman et al., 1993). Because related family members necessarily belong to the same population stratum, using relatives as controls automatically ensures protection against the effects of stratification.

Family-based association methods are by no means a panacea for complex trait gene mapping studies however. Families are often more difficult and more expensive to collect, especially for late-onset disorders where parents are unlikely to be available. In the absence of stratification, the simple case-control design is more powerful than the TDT: although a case-control pair and a TDT trio provide similar amounts of information, the case-control pair requires only two genotypes whereas the TDT requires three.

Recently, a different approach to population stratification has emerged: to use individuals' genetic backgrounds to detect stratification within a sample. If stratification is detected, genetic background data can be used as an index of ethnic grouping; tests of association robust to stratification can then be constructed, taking the stratification into account.

Signatures of stratification

A stratified sample will display certain characteristic 'signatures', both at single loci and also across unlinked loci: recent genetic-background methods detect stratification by looking for evidence of these signatures. At a single locus, stratification induces a non-independence between maternal and paternal alleles, i.e. Hardy-Weinberg disequilibrium (HWD). Across unlinked loci, stratification can induce a similar non-

independence of alleles, i.e. linkage disequilibrium (LD). For example, a single locus with alleles A_1 and A_2 occurring at frequencies p and q will, under HWE, have expected frequencies p^2, $2pq$ and q^2 for the three possible genotypes A_1A_1, A_1A_2 and A_2A_2, respectively. Considering two discrete subpopulations, P_1 and P_2, which differ greatly in allele frequency, the impact on HWE in the stratified sample $P_1 + P_2$ (assuming 50:50 admixture) is demonstrated below:

	Subpopulation		
	P_1	P_2	$P_1 + P_2$
A_1	0.1	0.9	0.5
A_2	0.9	0.1	0.5
A_1A_1	0.01	0.81	0.41 (0.25)
A_1A_2	0.18	0.18	0.18 (0.50)
A_2A_2	0.81	0.01	0.41 (0.25)

Assuming HWE within subpopulation, the expected genotype frequencies are tabulated for the two subpopulations (e.g. A_1A_1 genotype in subpopulation P_1 is 0.01). The allele and genotype frequencies in $P_1 + P_2$ are simply the average of those for P_1 and P_2. This creates a deviation from the expected HWE genotype frequencies in the stratified sample – the expected HWE genotype frequencies based on the allele frequencies in the admixed group are shown in parentheses. The typical "loss of heterozygosity" effect is illustrated here – only 18% instead of the expected 50% of the sample are heterozygous.

Although deviation from HWE at a single locus can be used as a weak test of population stratification, by itself it is indicative of several phenomena: (1) random sampling error (2) assortative mating (3) very high mutation rate (4) selection effects (5) unequal transmission ratios from parents to offspring (6) genotyping error (7) ascertainment effects (e.g. at the trait locus for cases in a case-control study) (8) and

population stratification. By itself, deviation from HWE at a single locus is not a specific signature of population stratification.

Expected heterozygosity can be used to quantify the magnitude of stratification within a sample. The expected average heterozygosity across random-mating subpopulations ($\bar{H}_S$) is compared to the expected heterozygosity in the total population (H_T). In our example above, $\bar{H}_S = 0.18$ and $H_T = 0.5$. Wright's fixation index is calculated $F_{ST} = (H_T - \bar{H}_S)/H_T$ and is a commonly used index of genetic distance between groups (n.b. in this context, 'genetic distance' refers to allelic frequency differences between groups, not genetic distance as previously defined in terms of recombination). In our example, $F_{ST} = 0.64$ which is a very large value. F_{ST} is always positive; 0 = panmixis (no subdivision, random mating occurring, no genetic divergence within the population); 1 = complete isolation (extreme subdivision). F_{ST} values up to 0.05 indicate negligible genetic differentiation whereas > 0.25 means very great genetic differentiation within the population analysed. For most European populations, one would expect values around 0.01–0.05; for the most divergent populations, one might expect values around 0.1–0.3 (Cavalli-Sforza et al., 1994).

Another signature of population stratification is association between alleles on unlinked loci. The following example contingency tables describe the association between alleles at two loci in 200 Scandinavians and 200 Spaniards, separately and then as a combined (i.e. stratified) sample:

Scandinavians

	B_1	B_2
A_1	160	160
A_2	40	40

Spaniards

	B_1	B_2
A_1	160	40
A_2	160	40

Combined

	B_1	B_2
A_1	320	200
A_2	200	80

When analysed separately, there is no association between the alleles at locus A and locus B (the χ^2_1 test of independence is 0 in both cases) for either Scandinavians or Spaniards. However, combining both samples gives a χ^2_1 of 7.81, which is significant

at the 5% level. This spurious association clearly would not be reflective of genetic distance – A and B could well be on different chromosomes.

1.7 Thesis outline

The motivation for this work is mirrored in the quotation below, taken from a recent NIH Request for Applications (RFA: MH-98-017) entitled "Quantitative methods to map genes for complex diseases":

> *Genetic factors contribute to virtually every human disease by conferring susceptibility or resistance, affecting the severity or progression of disease, and interacting with environmental factors that modify disease course and expression... Current analytic methods have been successfully applied to map Mendelian disease genes, but are not well suited for the genetic analysis of complex human diseases. Human geneticists are now beginning to explore a new genetic frontier, driven by the inconvenient reality that most diseases of medical relevance have irregular familial patterns and lack a simple one-to-one correspondence between genotype and phenotype.*

The majority of genetically-influenced public health concerns (traits and diseases such as depression & anxiety, hypertension, obesity and type II diabetes) can not easily be accounted for by single major locus models. Attempts to find the multiple genes of small effect that contribute to these traits have been plagued by low statistical power. Improving study design and optimising analytic tools is therefore a necessary next step: the use of selected sampling strategies to increase efficiency in QTL linkage and association studies is one route to this goal.

Mapping complex trait genetic architecture is likely to present challenges beyond small effect size: there will be no "one-to-one correspondence". Animal model studies increasingly reveal the significance of genetic background effects, pointing to the need to consider epistasis. Quantitative genetic studies are finding that environmental factors can often moderate the expression of genetic effects, implicating gene-by-

environment interaction. Additionally, different loci may be involved across ethnic strata, or polymorphisms may be differentially frequent. Considering these kinds of complex effects, in both unselected and selected samples therefore seems a worthwhile endeavour.

The Chapters following this Introduction are as follows:-

Part I : Sample selection

2. Selection for linkage

A novel method for selecting optimally informative sibships of any size for QTL linkage analysis is presented. The method allocates a quantitative index of potential informativeness to each sibship on the basis of observed trait scores and an assumed true QTL model. Any sample of phenotypically-screened sibships can therefore be easily ranked-ordered for selective genotyping.

3. Selection for association

The same strategy as described above for linkage can be applied to selecting sibships for family-based association analysis; furthermore, an approach to analysis is developed to provide a robust test of family association in selected samples. Several miscellaneous issues are considered at the end of the Chapter, including threshold selection for DNA pooling designs for quantitative traits.

Part II : Complex effects

4. Gene-by-environment interaction

This Chapter presents a basic model of $G \times E$, in the context of the twin design. Various aspects of the ability to detect different types of gene-by-environment interaction are investigated, as well as the consequences of not properly modelling interaction.

5. Epistasis

An extension to the variance-components QTL linkage model to incorporate epistasis between two loci is presented. Under a number of epistatic models, the power to detect the epistatic and main effects of two loci is calculated, providing an insight into the utility of multi-locus linkage approaches.

6. Population stratification

A method which approaches stratification as a latent class analysis (LCA) problem, similar to Satten et al. (2001), is developed and explored in this Chapter.

Part III: Sample selection and complex effects

7. Selection and gene–environment interaction

This Chapter considers two ideas: 1) incorporating knowledge of environmental interaction effects in order to increase the efficiency of selection and analysis for QTL linkage and 2) extending the conditional QTL linkage and association models, which are robust in selected samples, to include environmental interactions.

8. Selection and epistasis

The results of Chapter 5 refer to unselected samples: this Chapter develops a method of two-locus linkage valid in selected samples, based on an extension of the Haseman-Elston model. Additionally, the incorporation of a second, modifier locus in the association model presented in Chapter 3 is considered.

9. Selection and population stratification

This Chapter investigates some issues arising in the application of the genetic background method described in Chapter 6 in selected samples.

Part I

Sample Selection

Chapter 2

Selection for linkage

This Chapter presents a method of sample selection for variance components quantitative trait loci linkage analysis using sibships. The method involves the calculation of a quantitative index of potential informativeness, that reflects for each sibship it's expected contribution to the likelihood-ratio test statistic. The efficiency of this method compares favourably to other methods of extreme sample selection, including proband selection and extreme discordant pair selection.

2.1 Introduction

Optimal sample selection is an attempt to remedy the major practical problem of low power for the detection of quantitative trait loci (QTL). Risch and Zhang (1995) stated that a QTL would have to account for as much as 50% of phenotypic variance to be detectable in typically-sized unselected samples. As genotyping is still sufficiently expensive to prohibit increasing sample size in order to raise power to a desirable level, a partial solution is to select sibships for genotyping on the basis of phenotypic trait scores—that is, to genotype only the potentially most informative observations.

Lander and Botstein (1989) demonstrated the utility of selected samples in the context of experimental mouse studies: by selecting only phenotypically extreme an-

imals of an F_2 or backcross population, almost 90% of the linkage information could be recovered by only 25% of the sample. Subsequently, selected sampling approaches were adopted in human, sib-pair studies. Using thresholds to define phenotypically extreme individuals, three types of sib pair tend to confer the most information for linkage: concordant high, concordant low and discordant pairs. The affected sib-pair (ASP) (Suarez et al., 1978) design, successfully employed for mapping rare qualitative disorders, samples pairs concordant for being above a phenotypic threshold, e.g. genotyping only sib pairs in which both sibs fall in the top 10%. Clearly, the percentage of sib pairs selected will be dependent on the sibling correlation—typically, the threshold will be either based on *a priori* definitions of 'caseness' or will be suitably adjusted to yield the desired number of sib pairs. Proband selection (PS) (Carey and Williamson, 1991) ascertains sib pairs in which at least one sib scores above a phenotypic threshold, allowing both concordant high and discordant sib pairs to be selected. Risch and Zhang (1995, 1996) showed that, under most circumstances, extreme discordant (ED) pairs are more informative than concordant pairs. Under a positive sibling correlation, the mapping between phenotypic and genotypic discordance is stronger than between phenotypic and genotypic concordance. The extreme discordant and concordant (EDAC) strategy (Gu et al., 1996) selects both discordant and concordant sib-pairs. Thresholds for concordant and discordant pairs can be adjusted according to knowledge of the genetic model, in order to provide more efficient selective sampling. For example, if the researcher knew in advance that the putative QTL had a rare increaser allele, then the threshold for concordant high pairs could be lowered to include more concordant high pairs in the sample (reasons for why such sib pairs are more informative under such conditions are discussed below). Dolan and Boomsma (1998), recognising that most researchers will in fact not have knowledge of the genetic model for a putative QTL, conducted a study to identify general recommendations for setting the thresholds in the EDAC strategy, averaging

optimal threshold estimates over a wide range of genetic models.

Eaves and Meyer (1994), realizing that power could be gained from adopting a truly dimensional approach, hinted at the power of selecting maximally dissimilar (MDis) pairs on the basis of their squared trait difference. Methods that assign a quantitative index of informativeness to each sib pair enable genotyping to progress from the most informative down—significant results may be obtained before all of the selected sample has been genotyped. A more recently proposed strategy of selecting sibships for linkage uses the Mahalanobis distance (MahD) as a quantitative index of a sibship's informativeness, based on its phenotypic extremity (Allison et al., 1999a). For the i^{th} sibship, $d_i = \mathbf{x_i'}\mathbf{\Sigma^{-1}}\mathbf{x_i}$ where $\mathbf{x_i}$ is a vector of mean-centred trait scores and $\mathbf{\Sigma}$ is the sibling covariance matrix. Therefore d measures how close each sibship is to the multidimensional sibship mean. Unlike the methods described above, this method generalises in a very straightforward manner for larger sibships.

No selection strategy will be optimal under all possible circumstances. By 'optimal' we mean that if, for example, 5% of the sample is selected, no other 5% of the sample will provide more statistical power to detect linkage than the optimal set. Risch and Zhang (1996) have shown that it is more efficient to sample sib pairs concordant for extreme high trait values if the trait-increasing allele is rare (i.e. close to 0), but to sample sib pairs concordant for extreme low trait values if the trait-increasing allele is common (i.e. close to 1). The current approach allows an assumed true genetic model for the QTL to be sensibly parameterised and a selection strategy developed which would be optimal if the model were true. Such a method also allows exploration of the relationship between the accuracy of the assumed true model specification and the optimality of selective sampling.

The current method is based around the maximum-likelihood variance components approach, described in the Introduction. For linkage analysis, the likelihood function critically depends on the parameterisation of the covariance matrix, in terms of various

genetic and environmental components of variance and allele sharing identical-by-descent (IBD).

2.1.1 Measure of informativeness

The proposed method calculates the potential informativeness of a sibship of any size for QTL linkage analysis conditional on observed trait scores, under an assumed true genetic model. In a sample, power to detect a QTL is determined by the noncentrality parameter (NCP). The NCP is the sum of independent contributions from all the sibships in a sample. The expected contribution of a sibship to the sample NCP is therefore an index of potential informativeness for that sibship.

In computing each sibship's expected contribution to the NCP, all possible genotypic configurations (GC) are enumerated. For each sibship, the linkage test statistic is calculated under all possible sibship GC $(t_1, t_2, ..., t_n)$, assuming that the model parameters have been correctly estimated at their true values. These values represent the test statistics that would be obtained given the sibship's trait scores. Each test statistic is weighted by the probability of it occurring (i.e. the probability P_i of the i^{th} GC, given the trait scores) and summed over all GC to produce an index of potential informativeness for that sibship ($\sum P_i t_i$). Table 2.1 represents the calculation of this index; the posterior probability of the i^{th} GC conditional on trait values is given by Bayes Theorem as

$$P_i = P(GC_i|\mathbf{x}) = \frac{P(GC_i)f(\mathbf{x}|GC_i)}{\sum_{j=1}^{n} P(GC_j)f(\mathbf{x}|GC_j)}.$$

The proposed method therefore embodies the advantages of assigning a quantitative index of informativeness to sibships of any size that will be optimal if the assumed genetic model is true. Unlike the Mahalanobis distance, this measure is directly comparable across sibships of different size—to say whether a given sib pair is more or

GC	$P(GC\|\text{trait, model})$	χ^2 Test statistic	$P(GC\|\text{trait, model}) \times \chi^2$ Test statistic
1	P_1	t_1	$P_1 t_1$
2	P_2	t_2	$P_2 t_2$
3	P_3	t_3	$P_3 t_3$
...	...	...	...
n	P_n	t_n	$P_n t_n$
			$\sum P_i t_i = NCP = E(\chi^2\|\text{trait, model})$

Table 2.1: Calculation of the index of potential sibship informativeness.

less informative than a given sib trio, for example.

2.2 Methods

The genetic model assumes a fully informative marker 0cM from a hypothetical di-allelic QTL, which has increaser allele frequency p and decreaser allele frequency $q = 1 - p$, dominance to additive genetic value ratio $z = d/a$ and proportion of phenotypic variance accounted for by the QTL x. We assume no recombination since the presence of recombination should have no effect on the relative ranking of sibships with respect to their expected NCP based on trait values. Given these three parameters (p, z, x), as well as the trait variance (σ_T^2) and trait sibling intraclass correlation (r), the critically important parameters of the model are calculated: genetic values and variance components.

2.2.1 Genetic values

Using Falconer's notation, genotypes AA, Aa and aa are assigned genetic values of $+a$, d and $-a$, where a represents the additive effect and d represents the dominance deviation. These are functions of allele frequency, dominance to additive genetic value ratio, trait variance and the proportion of variance accounted for by the QTL, such that

$$a = \sqrt{\frac{x\sigma_T^2}{(2pq(1 + z(q - p))^2 + (2pqz)^2}},$$

and $d = za$. Additionally, these values are mean-centred such that the expected genetic population mean is zero, by expressing the values as deviations from $(a(p - q) + 2pqd)$.

2.2.2 Variance components

Total trait variance σ^2_T is decomposed into four orthogonal components: variance due to additive genetic effects at the QTL (σ^2_A), variance due to dominance genetic effects at the QTL (σ^2_D), variance due to residual polygenic genetic and shared environmental effects (σ^2_S) and variance due to residual polygenic genetic and nonshared environmental effects (σ^2_N).

Variance components are functions of genetic values and allele frequency: $\sigma^2_A = 2pq(a + d(p-q))^2$, $\sigma^2_D = (2pqd)^2$, $\sigma^2_S = r\sigma^2_T - \frac{\sigma^2_A}{2} - \frac{\sigma^2_D}{4}$ and $\sigma^2_N = (1-r)\sigma^2_T - \frac{\sigma^2_A}{2} - \frac{3\sigma^2_D}{4}$. The method of specification of genetic values ensures that σ^2_A, σ^2_D, σ^2_S and σ^2_N sum to equal σ^2_T. Note that σ^2_S is a function of sibling correlation once the shared effects of the QTL variance have been removed. Likewise, σ^2_N represents nonshared variance after the nonshared effects of the QTL have been removed.

2.2.3 Sibship genotypic configurations

The enumeration of all possible genotypic configurations is based upon parental mating types and inheritance vectors, implied by sibship size and number of alleles at the QTL. For a diallelic locus, the number of mating types is 2^4, or 16. The relative frequency of each type depends on allele frequencies (as well as population structure, e.g. the assumption of random mating). Inheritance vectors specify the identity-by-descent (IBD) status of any pair of siblings within a sibship. Table 2.2 illustrates the construction of the IBD values associated with each inheritance vector for trios. If parental mating type is coded as 12×34, then a sib's genotype can be written as 13, 14, 23 or 24. Therefore, a sib pair of 13 & 13 share 2 alleles IBD; a sib pair of 14 &

	Sib a		*Sib b*		*Sib c*		*IBD*			P
	pat	mat	pat	mat	pat	mat	ab	ac	bc	
Inheritance Vector										
1	1	3	1	3	1	3	2	2	2	$\frac{1}{64}$
2	1	3	1	3	1	4	2	1	1	$\frac{1}{64}$
3	1	3	1	3	2	3	2	1	1	$\frac{1}{64}$
4	1	3	1	3	2	4	2	0	0	$\frac{1}{64}$
...	...	...	...	...	...	...	...	...	...	...
64	2	4	2	4	2	4	2	2	2	$\frac{1}{64}$

Table 2.2: Inheritance vectors and identity-by-descent structure.

24 share only 1 allele IBD; a sib pair 13 & 24 share 0 alleles IBD. As mentioned, IBD allele sharing refers to sharing between pairs; for a sibship of size s, there will be 2^{2s} possible inheritance vectors, each of which expresses an IBD matrix of pairwise IBD values[1].

For a diallelic locus in sibships of size s, parental mating types and inheritance vectors combine to form 2^{4+2s} possible sibship genotypes. Each genotypic configuration therefore has three associated components: an associated IBD matrix of dimension $s \times s$ in which each element is either 0, 1 or 2; an associated genotype vector of dimension s, in which each element is either AA, Aa or aa; and an associated probability $P(GC) = P(\text{Parental Mating Type}) \times \text{P(Inheritance Vector)}$.

The IBD matrix generates Σ, the expected covariance matrix for each GC, because each IBD status (0, 1 or 2) has an associated expectation for sib covariance in terms of σ_A^2 and σ_D^2. In a similar fashion, the sibship genotype vector generates μ, the vector of predicted means. This is because each element of the sibship genotype has an associated genetic value, which can be partitioned into additive and dominance genetic effects.

[1] Although there are 2^{2s} configurations of inheritance vector for sibships size s, there are a smaller number of unique configurations. For trios, there are only 4 unique IBD configurations; for quads there are 8.

2.2.4 Expected NCP for linkage

For linkage, assuming zero-centred data, all QTL effects are modelled in the covariance matrix[2]. The parameters estimated are σ^2_A for additive effects and σ^2_D for dominance effects. Dominance is included in the model, though this may reduce power in small samples if the dominance variance is small. For pairs, the test for linkage involves parameterising $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma}_{\mathbf{L}} = \begin{bmatrix} \sigma^2_A + \sigma^2_D + \sigma^2_S + \sigma^2_N & \pi\sigma^2_A + z_2\sigma^2_D + \sigma^2_S \\ \pi\sigma^2_A + z_2\sigma^2_D + \sigma^2_S & \sigma^2_A + \sigma^2_D + \sigma^2_S + \sigma^2_N \end{bmatrix},$$

where π represents the proportion of alleles IBD and z_2 represents the probability of complete allele sharing IBD (thereby modelling dominance). Assuming complete marker information, if z_i represents the probability of sharing i alleles IBD, then $\pi = z_2 + z_1/2$. For larger sibships, the covariance matrix of dimension $s \times s$ is parameterised similarly, with the appropriate pairwise parameters in each off-diagonal element.

Under the null hypothesis of no linkage between the test locus and the QTL, the covariance matrix is

$$\boldsymbol{\Sigma}_{\mathbf{N}} = \begin{bmatrix} \sigma^2_A + \sigma^2_D + \sigma^2_S + \sigma^2_N & \frac{\sigma^2_A}{2} + \frac{\sigma^2_D}{4} + \sigma^2_S \\ \frac{\sigma^2_A}{2} + \frac{\sigma^2_D}{4} + \sigma^2_S & \sigma^2_A + \sigma^2_D + \sigma^2_S + \sigma^2_N \end{bmatrix}.$$

where the IBD probabilities at the QTL are independent of IBD status at the test locus, and so assume their prior probabilities of 1/4, 1/2 or 1/4 for sharing 0, 1 or 2 alleles respectively.

The likelihood under the alternative hypothesis of linkage varies depending on GC (critically, for linkage, only the inheritance vector determines the likelihood); the likelihood under the null is fixed across all GC. Following the procedure outlined in

[2]For association, the effects of the QTL are modelled in the means vector: see Chapter 3.

Table 2.1, sibship informativeness is given by

$$E(\chi^2|\mathbf{x}, model) = \sum_{i=1}^{GC} 2\left[\ln L_{Li} - \ln L_N\right] P(GC_i|\mathbf{x}).$$

The likelihood function is defined as in standard variance components models for linkage analysis (e.g. Fulker et al., 1999), but with an adjustment developed for the analysis of selected samples (Sham et al., 2000a). This adjustment, described in the Introduction, is based on conditioning on the observed trait values of the sibship, and makes the test robust in selected samples.

2.3 Implementation

A computer program to implement this method has been developed, hereafter referred to as SEL (**SE**lection for **L**inkage). As well as processing simulated data, SEL has been used to select sibships for genotyping in the GENESiS Study, a large community-based QTL study, which plans to select 600 optimally informative sibships from a sample of approximately 10,000 phenotypically screened sibships (Sham et al., 2001).

SEL is designed to read sibship trait scores for sibship sizes of 2 or more; the number of sibships is unrestricted[3]. As mentioned, data from sibships of variable size can be analysed together and the measure of potential informativeness is comparable across different sibship sizes (of course, in terms of efficiency, the informativeness of a sibship must be viewed with respect to sibship size, i.e. the number of genotypes that would be required). SEL can either calculate the mean, variance and intraclass sibling correlation from the sample or accept fixed values for these statistics (i.e. when pre-selected data come from a known population). SEL was written in Delphi to run under MS Windows and is freely available[4].

[3]Processing time increases exponentially with average sibship size but linearly with number of sibships.

[4]SEL can be downloaded from `http://statgen.iop.kcl.ac.uk/sel/`

2.4 Application to simulated data

2.4.1 Trait score simulation

Uniform data grids were generated in order to evaluate how information content varies with trait value. Two uniform data grids were generated, one for sib pairs and one for quads. The grid for pairs consisted of points at intervals of 0.2 across the range -4 to 4, such that the first pair was (-4,-4), the second pair (-4,-3.8) and so on up to the last pair (4,4). The grid therefore consisted of 41^2 (1681) points. A data grid for quads was also generated, using a less fine resolution of 0.5, giving 13^4 or 28561 points.

Datasets were also generated where the trait distribution contained a QTL effect superimposed on multivariate normal residuals. Datasets contained either 10,000 pairs, trios or quads. Eight sets were generated under different genetic models, as shown in Table 2.3. In practice, as it is unlikely the genetic model of a putative QTL will be known, it is important to assess the efficacy of selection when a 'base model' of equal allele frequencies and no dominance is specified instead of the true model. This was assessed for a range of true models: Models 1 and 2 represent unequal allele frequencies; Models 3 and 4 represent rare dominant disease genes; Model 5 represents a dominant gene under equal allele frequencies; Models 6 and 7 represent rare recessive loci.

For all models $\sigma_A^2 + \sigma_D^2$ was fixed to 0.1; σ_S^2 and σ_N^2 were fixed to 0.2 and 0.7 respectively, giving $\sigma_A^2 + \sigma_D^2 + \sigma_S^2 + \sigma_N^2 = 1$. In Models 3-7, which incorporate dominance, a and d were set to be equal (i.e. $d/a = 1$).

2.4.2 Selection for sibling pairs

The uniform data grids for pairs and quads were used to generate contour plots to describe the informativeness of different types of sibship under different genetic models. Figure 2.1 plots the sib-pair noncentrality parameter for linkage, as generated

Model	σ^2_A	σ^2_D	r	G(AA)	G(Aa)	G(aa)
Base $p = 0.50$; $a = 0.447$; $d = 0$	0.1	0	0.25	0.447	0	-0.447
Model 1 $p = 0.10$; $a = 0.745$; $d = 0$	0.1	0	0.25	1.341	0.596	-0.149
Model 2 $p = 0.25$; $a = 0.517$; $d = 0$	0.1	0	0.25	0.776	0.259	-0.259
Model 3 $p = 0.10$; $a = 0.404$; $d = 0.404$	0.095	0.005	0.249	0.654	0.654	-0.154
Model 4 $p = 0.25$; $a = 0.318$; $d = 0.318$	0.085	0.015	0.246	0.358	0.358	-0.278
Model 5 $p = 0.50$; $a = 0.365$; $d = 0.365$	0.067	0.033	0.242	0.183	0.183	-0.548
Model 6 $p = 0.75$; $a = 0.652$; $d = 0.652$	0.040	0.060	0.235	0.082	0.082	-1.222
Model 7 $p = 0.90$; $a = 1.59$; $d = 1.59$	0.018	0.082	0.230	0.032	0.032	-3.148

Table 2.3: Properties of simulated QTLs: for pairs, trios and quads. σ^2_S and σ^2_N are fixed to 0.2 and 0.7 respectively; r is the sib trait correlation; G(AA), G(Aa) and G(aa) are the genetic values for the three genotypes.

by SEL assuming the base model (z axis), as a function of sib-pair trait scores (x and y axes). The trait mean, variance and covariance are fixed to 0, 1 and 0.25 respectively. The increased informativeness of discordant sib pairs is clearly demonstrated. The extreme corners of the plot represent pairs where one sib is 4 standard deviations above the mean and one sib is 4 standard deviations below the mean—this is extremely unlikely to occur under bivariate normality with a positive sib correlation, of course. Note that most of the surface is relatively flat at a small value, which indicates that the majority of sib pairs are not particularly informative for linkage: this is why linkage analysis on unselected samples is so inefficient.

Figure 2.2 depicts the 5% selected sample when the base model (equal allele fre-

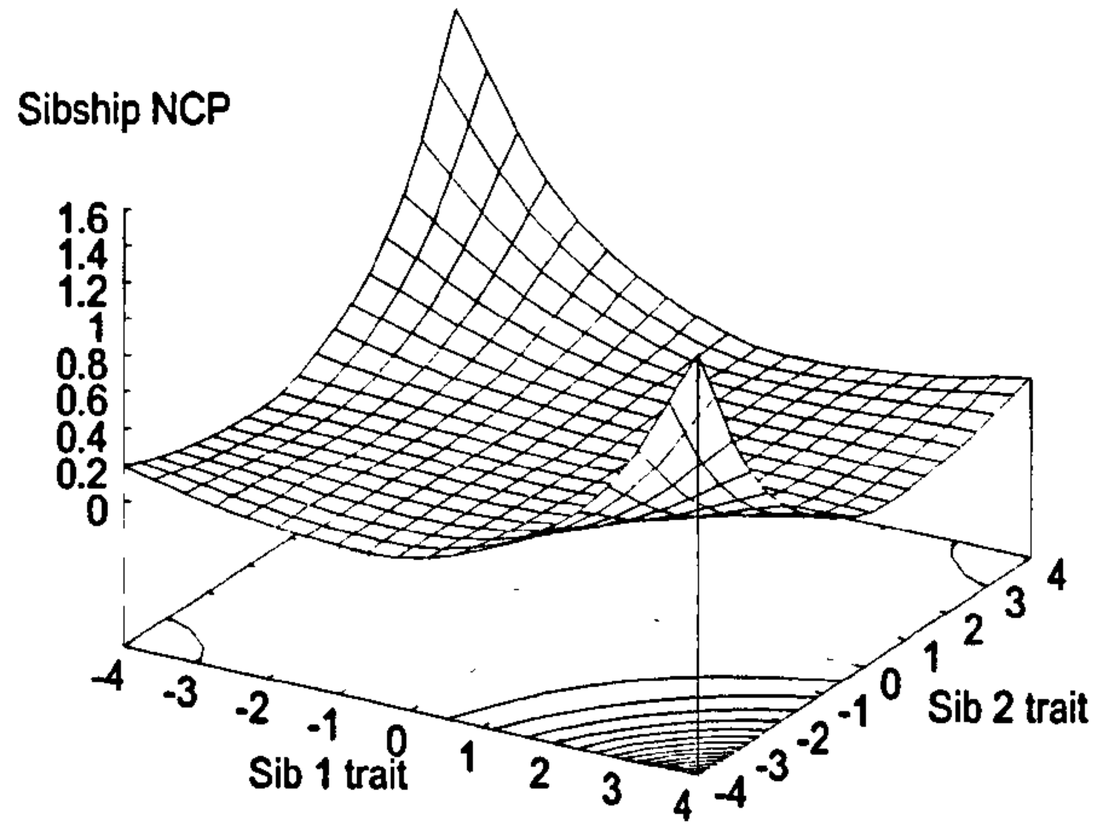


Figure 2.1: Contour plot of NCPs for sib pairs when the base model is true.

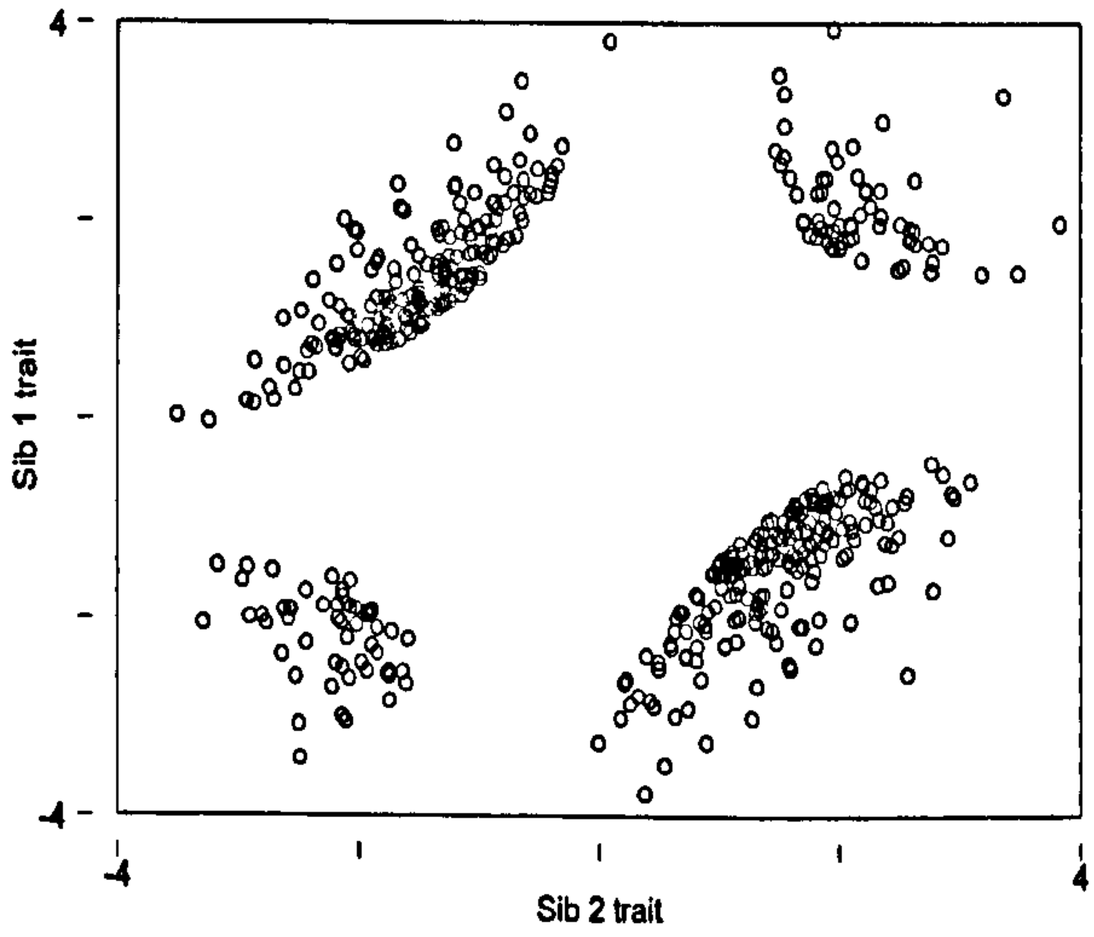


Figure 2.2: Scatter plot of the most informative 5% sib pairs when the base model is true.

quencies and no dominance) is true. Prior to selection, the sib-pair distribution is approximately bivariate normal with a sib correlation of 0.25. The Figure represents only the most informative 5% for linkage, as determined by the selection program. As can be seen, there is a preponderance of 'discordant' sib pairs in the selected sample.

Sib-pair informativeness as a function of sib-pair trait scores varies under different genetic models. In general, deviation from the base model in terms of dominance genetic variance and/or unequal allele frequencies results in the increased informativeness of one concordant quadrant complemented by the decreased informativeness

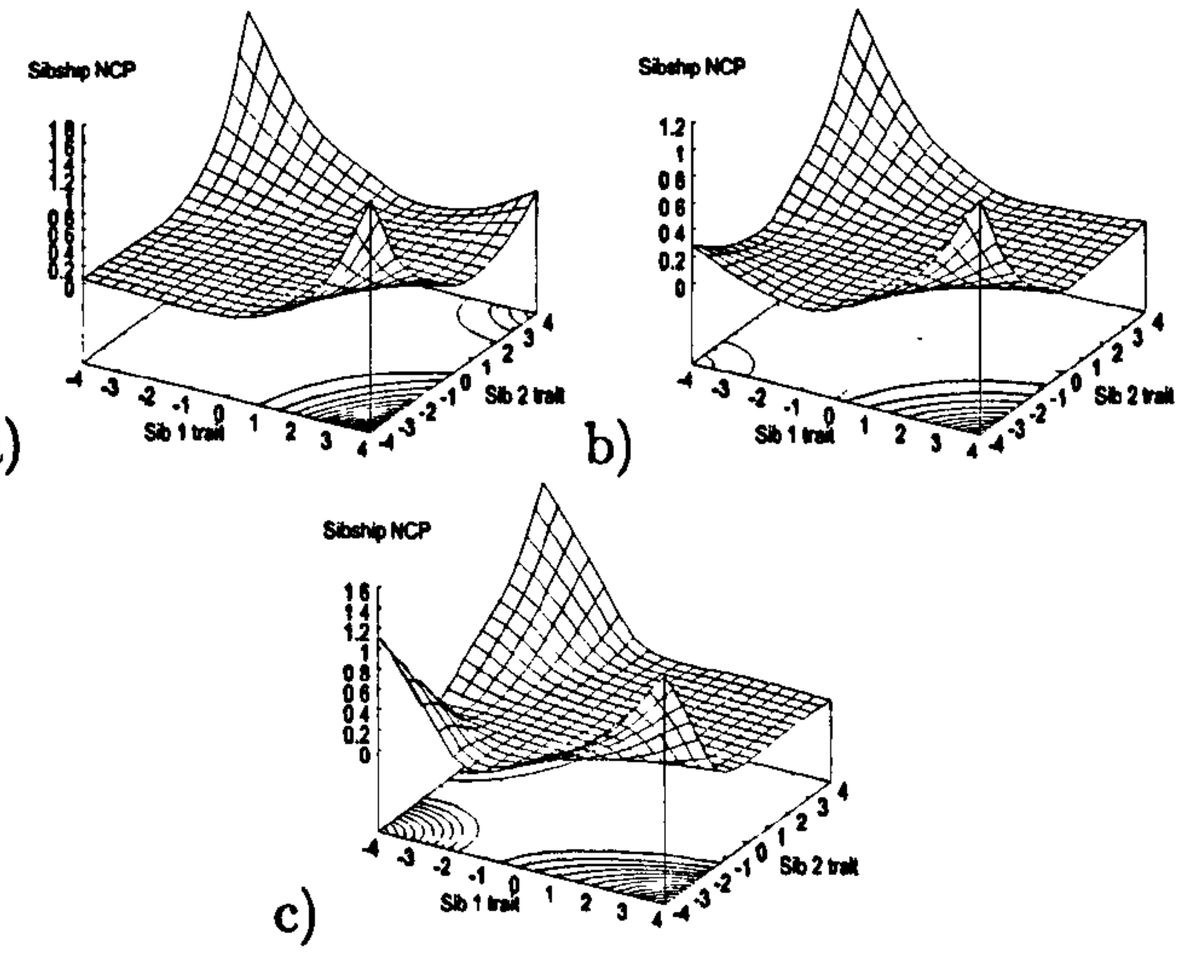


Figure 2.3: Contour plots demonstrating the effects of unequal allele frequency and dominance on selection: a) unequal allele frequency (p=0.1); b) dominance ($d : a$=1); c) rare recessive (p=0.9; $d : a$=1).

of the opposing concordant quadrant. To be precise, the end of the distribution associated with the less common allele, or the end of the distribution associated with the recessive allele, will be more informative. The effect of unequal allele frequencies tends to outweigh the effect of dominance. The first plot in Figure 2.3 illustrates sib-pair informativeness when the QTL has a relatively rare increaser allele ($p = 0.1$). The second panel represents a dominant QTL ($d/a = 1$). The third panel represents a rare recessive protective gene (i.e. a common dominant disease gene) ($p = 0.9; d/a = 1$),

Figure 2.4 demonstrates the effect of the residual sib correlation. In general, the greater the residual correlation, the greater the relative informativeness of discordant pairs. This can be conceived of in terms of the higher probability of phenotypically discordant sibs actually being genotypically discordant at the QTL under a higher residual correlation. Also note the marked difference in scale between the two figures: the highest sibship NCP under the low sib correlation is approximately 0.9 whereas under the higher sib correlation it is approximately 6.

The proportion of phenotypic variance accounted for by the QTL has no bearing on selection. Although this parameter will massively influence the power to detect the

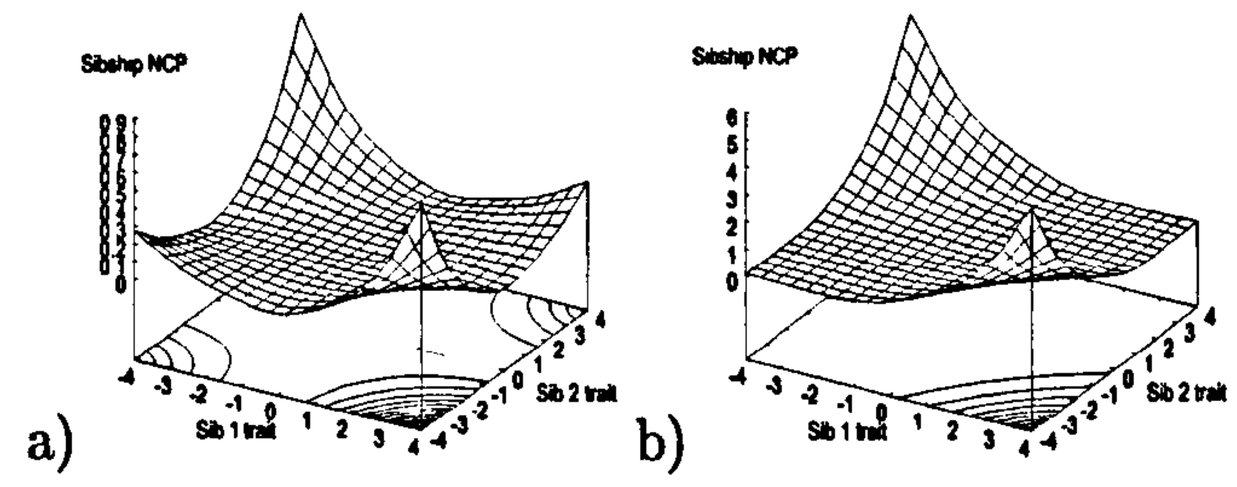


Figure 2.4: Contour plots demonstrating the effects of the residual sibling correlation on selection: a) low sib correlation (r=0.1) b) high sib correlation (r=0.5).

QTL, it does not effect the rank ordering of sibship informativeness, i.e. the sibship most informative for a small QTL will still be the most informative for a large QTL and vice versa.

2.5 Comparison with other selection strategies

It is difficult to generalise about precise selection criteria for complex traits, because they are a function of the genetic model which will not be known at selection. This suggests that 'mixed' approaches to selection may be preferable, i.e. including some sib pairs from all four quadrants. In the terminology of SEL, this corresponds to assuming the base model to be true. We are now in a position to investigate the effects of these assumptions.

SEL was run twice for each simulated dataset: under the true model (the model used to simulate that dataset) and under the base model (equal allele frequencies and no dominance). Under the true model, the sibship NCPs from the 5% most informative sibships were summed: this figure, F, represents the highest possible amount of linkage information that can be recovered from selecting 5% of the sample. In contrast, the second run, which assumes the base model to be true, will not necessarily select the same sibships as the 5% most informative. If S is the sum of NCPs *generated under true model* for the 5% most informative *selected under the base model*, then S/F represents the effect of mis-specifying the model. That is, S/F is the proportion

PAIRS	SEL^T	ASP	PS	ED	$EDAC^1$	$EDAC^2$	MDis	MahD	SEL^B
Base	15.82	24	33	77	84	81	89	81	100
1	17.09	50	40	68	82	81	78	80	96
2	15.45	50	32	74	82	80	87	80	99
3	16.88	48	38	69	83	82	81	80	96
4	15.76	36	25	80	81	79	90	78	100
5	18.39	26	15	76	81	78	87	78	98
6	27.64	13	4	56	71	69	73	83	91
7	43.16	1	5	25	69	62	52	98	95

Table 2.4: Results of simulations for pairs; SEL^T=SEL under true model; other values represent the percentage of information retained: ASP=Affected sib pairs; PS=Proband selection; ED=Extreme discordant; $EDAC^1$=Extreme discordant and concordant; $EDAC^2$=EDAC employing cut-offs recommended by Dolan & Boomsma (1998); MDis= Maximally dissimilar; MahD=Mahalanobis distance; SEL^B=SEL under base model.

of linkage information we would expect if we assume the base model versus if we knew the true model: this represents the 'optimality' of the selection scheme.

Table 2.4 represents these simulations for sib pairs; additionally, the efficiency of other sib-pair selection strategies under the different models are assessed. The first two columns describe the true genetic model. The column SEL^T gives the sibship NCP summed over the 500 most informative sibships (i.e. 5% selected from 10,000). As mentioned above, this figure represents the maximum amount of linkage information that 5% of that sample can contain.

It is possible to assess the efficiency of other methods of selection, using the NCPs generated under the true model as the metric, in the same manner as the base model comparison described above. The percentages in Table 2.4 therefore represent the extent to which the different methods are sub-optimal, under different genetic models.

Selecting affected sib pairs (ASP), the thresholds are adjusted such that 5% of sib pairs are selected. As can be seen, selecting concordant high sib pairs is not a good strategy: it performs moderately well when the increaser allele is rare; it performs abysmally when the increaser allele is common. Only in the case of a rare recessive disease gene (in our terminology, $p = 0.1$ & $d/a = -1$) would it be nearly optimal to select concordant high sib pairs. Proband selection (PS) is also inefficient compared

to other methods. Because the score of the second sibling is not constrained, most scores will be near the mean—giving very little power. The utility of proband selection is evident as a design issue: all other methods assume nonrandom selection from a large *randomly selected* population. Clinics and hospitals are clearly a rich source of probands who can be ascertained independent of their co-sib's status.

As has been previously demonstrated, discordant pairs are in general more powerful than concordant pairs. This is reflected in the higher average informativeness of ED selection under the different models. However, the lack of concordant sib pairs entails low power to detect rare recessives. EDAC is a more efficient sampling strategy than selecting only discordant pairs. Because concordant low pairs are sampled, for rare recessives over two-thirds of the linkage information contained in the optimal 5% of the sample is recovered. Thresholds were selected such that the selected sib pairs were comprised of 2.5% concordant (half concordant high, half concordant low) and 2.5% discordant. These are arbitrarily selected thresholds: as mentioned, Dolan and Boomsma (1998) calculated general recommendations for threshold values in the EDAC method. When selecting 5% from a random sample of 10,000 pairs, the recommendations suggest that pairs are concordant when both sibs are in the top or bottom 7.6% of the phenotypic distribution; pairs are discordant when one sib is in the top 17% and the other sib is in the bottom 17%. As is evident in Table 2.4 however, these general recommendations actually result in consistently *less* efficient selection than fixing each quadrant to be of equal size, presumably because fixing equal proportions is actually sensitive to the sib correlation whereas the general recommendations are not.

A more powerful method of selection relies on ranking pairs in terms of their squared trait difference and selecting the maximally dissimilar 5% (MDis)[5]. This method is more powerful than the ED method. This essentially implies that sib pairs

[5]Eaves & Meyer originally suggested selective genotyping of maximally similar sib pairs.

where one sib is very extreme but the other is relatively near the mean are often more informative than sib pairs consisting of a just-above-threshold high sib and a just-below-threshold low sib.

Selection using Mahalanobis distances (MahD) fixes the trait mean, trait variance and trait sib covariance to the values used to simulate the data, namely $\bar{x} = 0$ and $\Sigma = \begin{bmatrix} 1 & 0.25 \\ 0.25 & 1 \end{bmatrix}$. Assuming bivariate normality and dependent upon the sib correlation, approximately equal proportions of sib pairs from the four quadrants are selected, like the EDAC method but unlike MDis. However, MahD selection will also not preclude the inclusion of the potentially very informative configuration where one sib is very extreme and the other sib scores near the mean, unlike the EDAC method but like MDis. As a result of this, it is interesting to note how well this method performs in the case of a rare recessive gene. Despite the facts that concordant-high sib pairs afford no information in this case and that this method selects as many concordant-high as concordant-low (assuming no skewness or heteroscedasticity) the method is so near to optimality (98%) because, under this model, so few sib pairs actually *will* be informative. In the case of a rare recessive as in Model 7, less than 5% of the sample will contain over 90% of the information for linkage (see 'Efficiency of selection' section below). In this case, the 5% most informative selected on the basis of MahD will comprise two types of pair: approximately a fifth to a quarter will be concordant-low and very informative; the rest will not be particularly informative, but *none of the unselected pairs* will be informative either.

The final column of Table 2.4 illustrates the very high efficiency of the selection method implemented in SEL assuming the base model. In all of the simulated models, assuming the base model to be true yields more than 90% of the information that would be obtained if the optimal 5% were selected with full knowledge of the genetic model. Model 6 is the least efficient under the base model—in this case, the unselected 95% *will* contain sib pairs informative for linkage, because the overall proportion of

	$\sum NCP$	$\sum NCP / \sum NCP_{pairs}$
Pairs	30.024	1
Trios	94.057	3.133
Quads	193.215	6.435

Table 2.5: Relative average informativeness of sib pairs, trios and quads under the base model.

sib pairs informative for linkage will be higher than in Model 7. In general, on the basis of these data we recommend the proposed novel method of sib-pair selection as the most nearly optimal under a wide range of uncertain genetic models.

2.5.1 Selection of larger sibships

Up until this point, all results have been confined to the selection of sib pairs. This is because the existing strategies are primarily confined to the selection of sib pairs. However, analytic work has demonstrated the increase in efficiency that comes from analysing larger sibships (Sham et al., 2000b). To address this issue, samples of 10,000 pairs, 10,000 trios and 10,000 quads were generated under the base model. An advantage of the current strategy is that an index of informativeness is assigned to a sibship irrespective of sibship size. Expected sibship NCPs can therefore be used to assess the relative average informativeness of pairs, trios and quads for unselected samples. Table 2.5 demonstrates that the average informativeness of these three sibship sizes increases approximately in the ratio of 1:3:6. This ratio represents the number of unique pairwise combinations possible in pairs, trios and quads respectively. Of course, genotyping a quad involves twice as many PCRs, but will on average yield approximately 6 times the information as compared to a sib pair.

Similarly as for sib pairs, a uniform trait score grid was constructed for quads. Figure 2.5 represents the informativeness of quads in a matrix of contour plots. Trait scores for the third and fourth siblings were categorised into a seven-by-seven grid; for each grid point the contour plot of sibship informativeness as a function of the

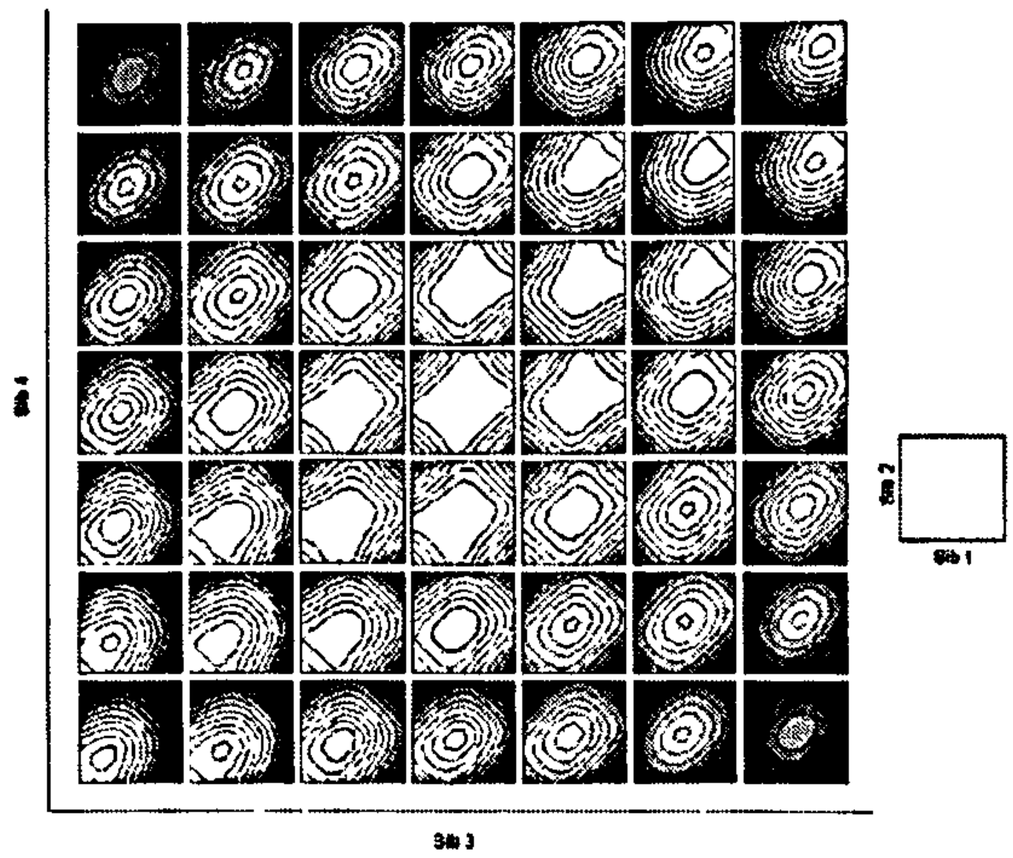


Figure 2.5: Matrix of contour plots to demonstrate sibship informativeness for quads when the base model is true; darker shades represent higher informativeness; trait scores for 3^{rd} and 4^{th} sibs banded into seven bins; small contour plots for 1^{st} and 2^{nd} sibs.

continuous scores for the first and second sibling was generated. The contour gradients represent the expected NCP for that sibship; the darkest bands represent the most informative sibships. For example, the bottom left corner of the bottom left plot represents four sibs all concordant low. In contrast, the top left corner of the bottom right plot represents scores that are low, high, high and low for sibs 1 to 4, respectively. Quads appear to embody the same principles of informativeness as pairs: that is, discordant sibs are generally more informative. There are surprisingly uninformative areas around the all-four-sibs concordant high and concordant low regions. Arguably, the increased power of larger sibships is a function of the increased probability of encountering genotypic and/or phenotypic discordance within a family.

Table 2.6 assesses the efficiency of SEL for larger sibships selected under the base model, in comparison with selection based on Mahalanobis distances (MahD). Relative to pairs, SEL performs slightly less well with larger sibships; the MahD method performs slightly better. In absolute terms, however, SEL is more efficient overall.

	TRIOS			QUADS		
Model	SEL^T	SEL^B	MahD	SEL^T	SEL^B	MahD
Base	36.99		85%	64.06		88%
1	45.83	94%	85%	80.24	94%	88%
2	39.05	99%	85%	61.65	98%	86%
3	42.72	95%	84%	78.16	95%	86%
4	36.07	100%	83%	65.04	100%	87%
5	40.58	97%	84%	73.83	97%	87%
6	63.66	88%	85%	135.70	89%	87%
7	138.79	95%	98%	255.47	96%	98%

Table 2.6: Results of simulations for trios and quads; SEL^T=SEL under true model; SEL^B=SEL under base model; MahD=Mahalanobis distance.

2.5.2 Efficiency of selection

Table 2.4 above represented the informativeness of a 5% sample selected under the base model relative to the informativeness of the optimal 5% which assumes knowledge of the QTL. These values for 5% groups selected under the base model can be re-expressed as proportions of the total amount of linkage information in the entire sample under the true model. That is, we can straightforwardly ask what proportion of all available linkage information would be obtained by selecting a n% sample assuming the base model. Figure 2.6 plots the relationship between the proportion of the sample selected and the proportion of linkage information recovered when the base model is true. If selection were random, one would expect a straight line at 45°, i.e if one selects 50% of the sample, then one expects 50% of the information. The extent to which the line is curved upwards represents the efficiency of selection, i.e. the extent to which $m > n$ if we select n% obtaining m% of the total information for linkage. Lines are drawn for pairs, trios and quads; the proportion of information is scaled to the total amount of linkage information available in the sample *for that sibship size* – it does not imply that pairs, trios and quads are equally informative when unselected (as has already been demonstrated to not be the case). In fact, the curve deviates more greatly for sib pairs than for trios; the curve for trios deviates more than the curve for

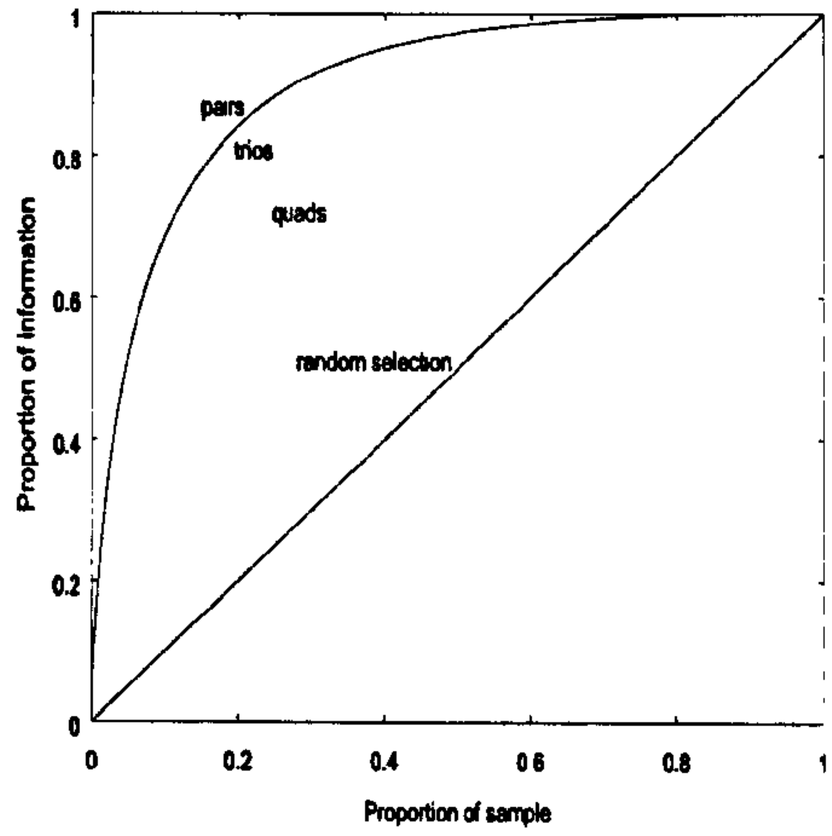


Figure 2.6: Efficiency of selection for different size sibships when the base model is true.

quads. This implies that selection has a greater impact on sib pairs than for trios and quads. This result is due to the fact that there is a higher proportion of almost totally uninformative sib pairs than sib trios or quads: those sib pairs who are IBD 1. In this instance, as expected gene-sharing does not deviate from the population mean, the expectations for the sib pair's trait scores also follow the expectations for the population mean. However, because it is not possible for all the pairs in a trio or quad to be IBD 1, the information in a sample of trios and quads is distributed more evenly across sibships. So, as the least informative trios and quads will be *relatively* more informative (i.e. considering sibship size) than the least informative pairs, this implies that selection will be more efficient for pairs.

The efficiency curves obtained when other genetic models are true (but selection still assumes the base model) are largely similar to the case under the base model. However, Figure 2.7 shows the curves for Model 7, a rare recessive QTL, which illustrates the point outlined above: that only a small percentage of the population will contain virtually all the linkage information for that QTL. Differences between pairs, trios and quads disappear in this case—the curves are virtually indistinguishable.

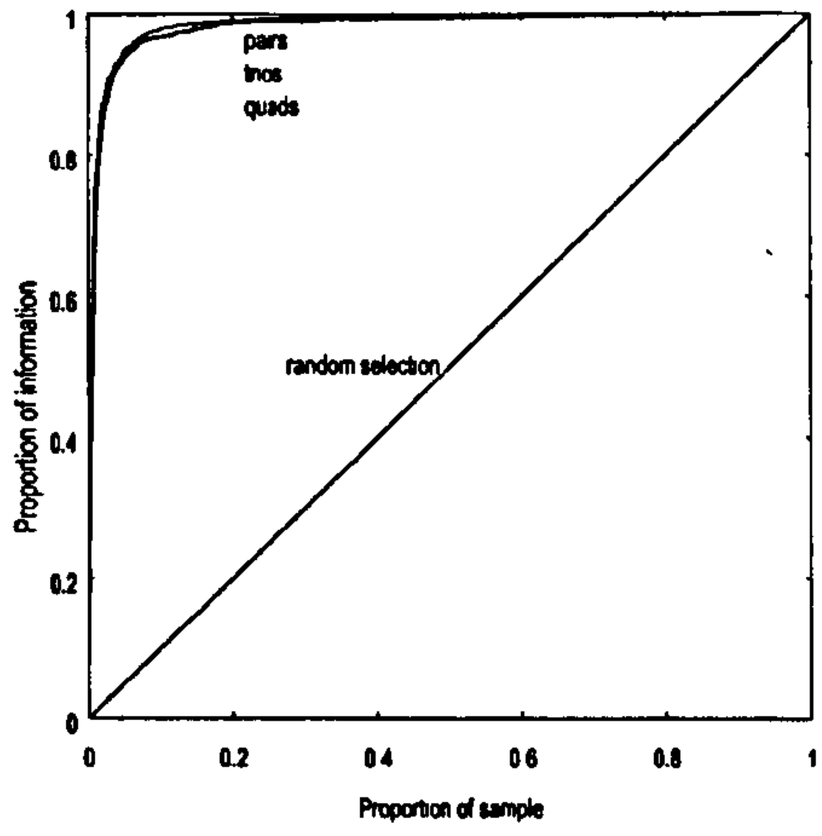


Figure 2.7: Efficiency of selection for different size sibships when the QTL is rare recessive.

2.5.3 Selecting optimal subsets of siblings from larger sibships

When selecting larger sibships, not all pairwise combinations within one sibship will be equally informative. That is, one or more sibs may be dropped to select a subset of the sibship for genotyping without significant loss in potential informativeness. If, for example, in a trio one sib scores very high, one scores very low and the other scores at the mean, it is highly likely that the pair consisting of the two extreme sibs would be more efficient to genotype than the whole trio (that is, if the pair's NCP is not less than two-thirds of the trio's). The program implements an option to output the informativeness of larger sibships conditional on each member being dropped one at a time. This information can be used to easily select the most efficient sib configuration from a sibship in order to improve efficiency.

2.6 Haseman-Elston linkage analysis

It is of interest to note that investigations along parallel lines, looking at the Haseman-Elston approach to linkage (Haseman and Elston, 1972), have revealed an approach to selection that is equivalent to the method proposed in this Chapter. Based on an

extended Haseman-Elston test (Sham and Purcell, 2001), the expected NCP for pair i is

$$\frac{q^4}{16}\left[\frac{(T_{i1}+T_{i2})^2}{(1+r)^2}-\frac{(T_{i1}-T_{i2})^2}{(1-r)^2}+\frac{4r}{1-r^2}\right]^2$$

assuming complete marker informativeness, where T_{i1} and T_{i2} are the standardised trait scores for the pair, r is the sibling correlation, and q^2 is the proportion of variance due to the QTL. This approach has also been extended to general pedigrees (Sham et al., 2002b).

2.7 Summary

This Chapter has presented a strategy for selecting sibships for genotyping on the basis of trait scores that is optimal when the genetic model is known and is more efficient than other strategies in the case when the model is unknown. This strategy has been implemented in the computer program SEL. The basic method of selection outlined here could be easily extended in several ways. Three possible extensions are summarised here: multivariate selection, the incorporation of family structures other than sibships and the modelling of assortative mating.

Selection could occur on the basis of multivariate trait data. Separate genetic values and variance components would need to be specified for each trait; additionally, the covariances between these components would have to be specified (e.g. residual polygenic genetic correlations). A potential problem is that more arbitrary decisions regarding the assumed true model would be required prior to selection. If these decisions are correct, then multivariate selection will indeed be more powerful than selection based on the best possible univariate composite of the same measures. However, if they are not correct, then multivariate selection will be potentially less powerful. Therefore, in the absence of good background knowledge regarding the aetiological architecture of the covariance between traits, selection on one composite measure is

arguably preferable, if the researcher wishes to incorporate information from different variables. Of course, analysis need not use the same dependent measure as selection, although this will tend to reduce power.

It is possible to extend the current method of selection to family structures other than sibships. Indeed, the current method is very similar to the method of expected lod scores proposed by Ott (1991) to measure linkage information content of pedigrees for classical linkage analysis. It is desirable to be able to select observations from such structures on the basis of their potential informativeness, if only to explore the theoretical impact of selection for such family groups. One obvious extension to the sib pair design is to allow the inclusion of half sibs. Under certain conditions (e.g. assumed true genetic model, sibling correlation, relative cost of phenotyping versus genotyping) it may prove beneficial to focus sample collection on obtaining larger unselected families rather than small selected sibships (Alcais and Abel, 2000), although, as mentioned, selecting the informative subsets from larger families may still be desirable.

Finally, the genetic model utilised in selection is a simple one which could be easily extended to include other genetic phenomena. For example, nonrandom mating could be modelled via the appropriate specification of the frequencies of parental mating types (e.g. $AA \times AA$ type might not necessarily occur at rate p^4 in the population). However, as with the multivariate extension, there is an argument in favour of keeping the number of parameters that need be specified to a minimum. Even if a trait shows evidence of nonrandom mating, it is conceivable that the majority of QTLs for that trait do not display such effects in any significant extent. The current results might not hold for certain oligogenic models which involve epistasis. Because the proposed method samples from all four corners of the bivariate distribution, however, it is likely to be more robust than other methods of selection which only sample from one extreme.

A potential limitation of the current method is its reliance on multivariate normality in the data. If trait data are skewed or contain outliers, certain sibships will receive unrepresentative measures of informativeness. Although such observations would bias results of analysis in any case, the more extreme the selection, the greater the proportion of extreme, artefactual outliers to genuine cases. Additionally, using discordant sib pairs may tend to increase the rate of half siblings due to non-paternity in the sample (Allison et al., 1998). The misclassification of half siblings as full siblings would inflate the false positive rate of linkage. It is therefore important to identify half siblings by checking the proportion of alleles IBD over a number of marker loci.

Another simplification in the current method is that we have assumed IBD information to be complete. This simplification is unlikely to have a major impact on selection for a multipoint linkage analysis using highly polymorphic markers, especially if parental genotypes are available. However, if parental genotypes are unavailable and marker information is far from complete, then larger sibships will have the further advantage over sib pairs that the multiple sibling genotypes will help to provide more accurate estimates of the IBD sharing between siblings.

As mentioned, the proposed selection method was developed within the framework of maximum-likelihood variance components QTL linkage analysis. Although this has not been empirically tested, the described method of selection should be nearly optimal irrespective of the method of analysis. Because maximum-likelihood estimates are asymptotically unbiased, the proposed selection method should be applicable to any robust method of analysis, such as Haseman-Elston regression analysis (in fact, this is implicitly demonstrated in Sham and Purcell (2001) and Sham et al. (2002b)).

The principles of the current method can be applied in other situations where it is necessary to select a subset of a sample for more extensive study. One such alternative scenario is the allelic association study, outlined in the next Chapter.

Chapter 3

Selection for association

As discussed in Chapter 2, linkage studies of quantitative traits in sibships are very inefficient if unselected samples are used. In this Chapter, the method outlined in Chapter 2 for calculating an index of expected informativeness for each sibship (that can be used to select only the most informative sibships for genotyping) is applied to variance components quantitative trait loci association analysis in sibships. Using selected samples improves the efficiency of association studies, although not to such an extent as for linkage studies. A 'conditioning-on-trait-values' approach to the analysis of selected samples is also presented, implemented in a computer program, along with simulation results. Finally, sample selection issues in three other scenarios are briefly considered: an approximation for the informativeness of selected samples; the power of two-stage designs; determining pool thresholds in DNA pooling designs with quantitative traits.

3.1 Background

Extreme sample selection has been applied to association as well as linkage designs (e.g. Petrill et al., 1997), in an attempt to increase power to detect genes of small effect for complex, quantitative phenotypes. The most basic selection procedure di-

chotomises a quantitative trait measured in a population of unrelated individuals into two groups of "cases" and "controls", reflecting high and low scorers. In this scenario, the most efficient design strategy does not involve dichotomising the entire population (e.g. a median split) since selecting groups only from the two extremes of the distribution increases the effective QTL effect (Van Gestel et al., 2000). The simplest form of analysis compares allele frequencies in the two groups. Although this approach loses the quantitative information available, this is not necessarily a concern in the context of DNA pooling, in which pool allele frequencies are compared between high and low groups. A section at the end of this Chapter briefly considers some issues in sample selection for DNA pooling studies.

Power analyses for this kind of threshold-defined case-control analysis can be readily performed using the Genetic Power Calculator (GPC) (Purcell et al., 2003). Consider the following simple example: an unselected sample of 1000 singletons and a QTL explaining 2% of the total variance (additive effects and equal allele frequency is assumed). The noncentrality parameter (NCP) for a quantitative association analysis is 20.2, which corresponds to 99% power at $\alpha = 0.05$ and 86% power at $\alpha = 0.001$. In contrast, if only 10% of the sample is selected, evenly from the high and low ends of the trait distribution (i.e. a threshold of 1.645 standard deviations from the mean) then the NCP for a case-control design is 8.41 (power 83% at $\alpha = 0.05$ and 35% at $\alpha = 0.001$). In this instance, sample selection and dichotomous analysis is relatively efficient, as it retains 42% of the information for association with only 10% of the sample.

However, dichotomising a continuous trait into high and low groups and performing a case-control analysis is not necessarily the best approach. This Chapter outlines an approach by which singletons (or sibships) can be rank-ordered by their potential informativeness for association; in addition, a robust method for analysing selected samples in a quantitative manner is presented, which is often more powerful. For the

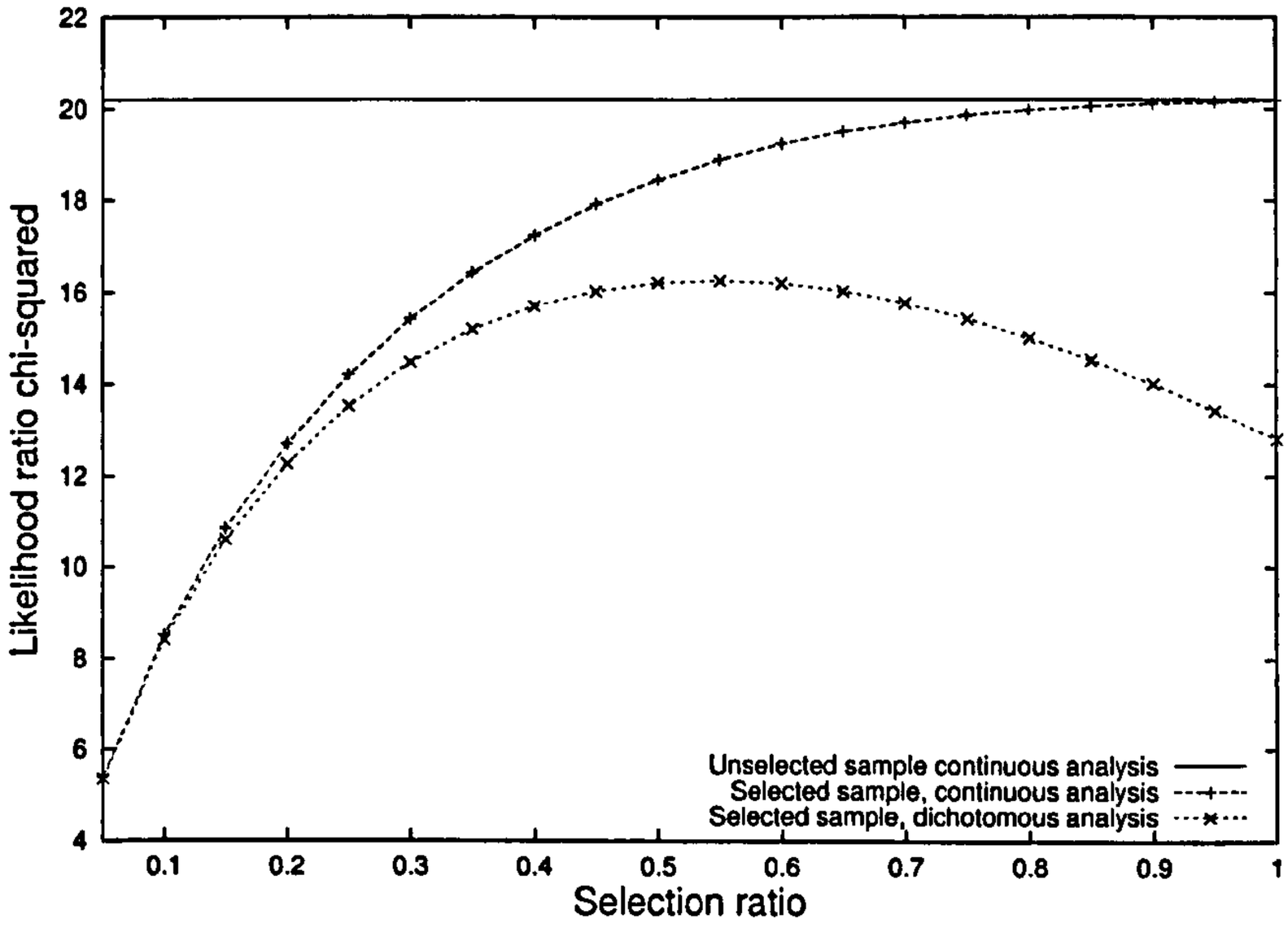


Figure 3.1: Comparison of quantitative and threshold-based "case-control" association analyses: singletons only.

simple case of singletons, as sample selection becomes more extreme, dichotomous and continuous analytic approaches become equivalent. This is because there is little within group variation – virtually all of the variation in the selected sample is captured by dichotomous group membership. However, as the proportion of the selected sample becomes larger, a continuous approach to analysis becomes more powerful, as illustrated in Figure 3.1.

The discrepancy between a threshold-based and a continuous approach to the analysis of selected samples becomes greater when considering larger sibships. Abecasis et al. (2001a) compared dichotomous and continuous family-based tests of association in selected samples: although under certain conditions when there is very little quantitative variation in the selected sample (e.g. all pairs selected concordant high) a dichotomous approach performs better, in the majority of cases a quantitative approach is preferred. The rest of this Chapter focuses on continuous analysis.

3.2 Fulker association model

The Fulker model (Fulker et al., 1999) offers an elegant approach to combining both linkage and association analyses for sibships within the same maximum-likelihood variance components framework. Furthermore, association is divided into orthogonal between and within sibship components, where the within component is unaffected by population stratification and admixture effects, and so provides the basis for a robust test of association (cf. Chapters 6 and 9 look at alternative approaches to population stratification and admixture).

The between (B) and within (W) components are based on the sibship genotypic mean and the intra-sibship differences in genotype, respectively. For a single diallelic marker locus (which is the QTL itself) Table 3.1 presents the partitioning of the additive effect into between and within components. The 'Additive effect' columns give the genotypic effect in terms of a, the additive genetic value for the locus. The next two columns, labelled 'Components' show the between and within components of association for each sibship genotype. That is, the between component represents the average genotypic effect, labelled a_b, whilst the within component is half the difference between the siblings' genotypic effects, labelled a_w. Because siblings will necessarily belong to the same population stratum, within sibship association cannot be due to stratification effects. Therefore, a_w should provide a more robust (albeit possibly less powerful) test of association than a_b (or a_b and a_w combined, called total association). The final two columns show how the genotypic effects are partitioned in terms of between and within components. Note that if $a_b = a_w$ (i.e. no stratification or admixture effects) then the scores reduce to the basic formulation in terms of a single a.

More formally, let $\mathbf{g}$ be the vector of genotypic scores for a sibship of size s. Under a total association model, the sibship expected mean vector is $\mathbf{g} = a\mathbf{A} + d\mathbf{D} - m$ where each element of $\mathbf{A}$ represents an individual's additive effect coded 1, 0 or -1

Genotype		Additive effect		Components		Partitioned effects	
Sib 1	Sib 2	Sib 1	Sib 2	Between	Within	Sib 1	Sib 2
1/1	1/1	a	a	a_b	0	a_b	a_b
1/1	1/2	a	0	$a_b/2$	$a_w/2$	$a_b/2+a_w/2$	$a_b/2-a_w/2$
1/1	2/2	a	$-a$	0	a_w	a_w	$-a_w$
1/2	1/1	0	a	$a_b/2$	$-a_w/2$	$a_b/2-a_w/2$	$a_b/2+a_w/2$
1/2	1/2	0	0	0	0	0	0
1/2	2/2	0	$-a$	$-a_b/2$	$a_2/2$	$-a_b/2+a_w/2$	$-a_b/2-a_w/2$
2/2	1/1	$-a$	a	0	$-a_w$	$-a_w$	a_w
2/2	1/2	$-a$	0	$-a_b/2$	$-a_w/2$	$-a_b/2-a_w/2$	$-a_b/2+a_w/2$
2/2	2/2	$-a$	$-a$	$-a_b$	0	$-a_b$	$-a_b$

Table 3.1: Partitioning of additive effects into between- and within-pair components (after Fulker et al., 1999).

corresponding to genotypes 1/1, 1/2 and 2/2 respectively; likewise, **D** represents dominance effects, coded 0, 1 or 0 respectively. The variables a and d are the free parameters to be estimated. The variable m represents the mean – this is subtracted to ensure the expected population mean is fixed to zero, and is calculated $a(p-q)+2pqd$. Two vectors can be calculated to correspond to the between and within components, for both additive and dominance effects. For example, for additive effects, the between effects vector $\mathbf{A_b}$ contains the sibship mean

$$[\mathbf{A_b}]_i = \frac{\sum_{j=1}^{s}[\mathbf{A}]_j}{s}$$

whilst the within effects vector $\mathbf{A_w}$ contains the individual deviations from the sibship mean

$$\mathbf{A_w} = \mathbf{A} - \mathbf{A_b}$$

Repeating this procedure for dominance effects, the full means model is

$$\mu = a_b\mathbf{A_b} + a_w\mathbf{A_w} + d_b\mathbf{D_b} + d_w\mathbf{D_w} - m$$

where $m = a_b(p-q)+2pqd_b$. The within components do not contribute to the overall

mean, as $E([\mathbf{A_w}]) = E([\mathbf{D_w}]) = 0$. The residual covariance matrix is simply

$$[\Sigma]_{ij} = \begin{cases} \sigma_S^2 + \sigma_N^2 & \text{for } i = j \\ \sigma_S^2 & \text{for } i \neq j \end{cases}$$

where σ_S^2 represents residual shared sibling variance and σ_N^2 represents residual non-shared sibling variance. By fixing or equating a_b, a_w, d_b and d_w under alternate and null models, likelihood ratio test statistics can be calculated representing various tests of association effect. For additive only models, the total test involves the models $\mathrm{H_A}(a_b = a_w)$ against $\mathrm{H_0}(a_b = a_w = 0)$. A robust test can be constructed in two ways: testing within association either in the presence of between association effects or not. That is, explicitly modelling between effects, the robust test compares $\mathrm{H_A}(a_b, a_w)$ against $\mathrm{H_0}(a_b, a_w = 0)$; otherwise comparing $\mathrm{H_A}(a_b = 0, a_w)$ against $\mathrm{H_0}(a_b = a_w = 0)$ also provides a robust test. The differences between these two formulations are explored in the simulation sections below. A test of stratification is $\mathrm{H_A}(a_b, a_w)$ against $\mathrm{H_0}(a_b = a_w)$. To model linkage simultaneously, a further term for the QTL variance is introduced into the covariance model which is dependent on IBD sharing at the test locus (see Chapter 2).

If the vector $\mathbf{x}$ contains the trait scores for a sibship, the log-likelihood of observing $\mathbf{x}$ conditional on the observed sibship genotype, assuming normality, is

$$\ln L(\mathbf{x}|\mathbf{g_o}) = -\frac{1}{2}\left[-\ln|\mathbf{\Sigma_i}| - (\mathbf{x} - \mu)'\mathbf{\Sigma_i}^{-1}(\mathbf{x} - \mu)\right]$$

and likelihood ratio tests are twice the difference in log-likelihood between null and alternate hypotheses.

P	M	A_b	D_b
1/1	1/1	1	0
1/1	1/2	0.5	0.5
1/1	2/2	0	1
1/2	1/1	0.5	0.5
1/2	1/2	0	0.5
1/2	2/2	-0.5	0.5
2/2	1/1	0	1
2/2	1/2	-0.5	0.5
2/2	2/2	-1	0

Table 3.2: Between-sibship scores for additive (A_b) and dominance (D_b) effects when paternal (P) and maternal (M) genotypes are available.

3.2.1 Parental genotypes

Parental genotypes, when available, can be used to construct the additive and dominance between-sibship vectors for that sibship. That is, conditional on parental genotypes, the expected values for $\mathbf{A_b}$ and $\mathbf{D_b}$ can be easily calculated by considering the 4 equally likely potential offspring genotypes (see Table 3.2). One important consequence of using parental genotypes is that singletons with parents genotyped are now informative for the robust within test of association (Abecasis et al., 2000). Otherwise, the within component is necessarily zero for singletons.

3.3 Conditional association test

As mentioned in the Introduction, and as for linkage analysis, standard tests of association are not necessarily robust or optimally powerful in selected samples. In this section, the conditioning-on-trait-values approach described in the case of QTL linkage analysis (Sham et al., 2000a) is extended to the case of QTL association analysis.

The standard association model considers the likelihood of the trait conditional on observed genotype, $L(\mathbf{x}|\mathbf{g_o})$. Alternatively, it is possible to consider the likelihood of the genotype, conditional on trait values, $L(\mathbf{g_o}|\mathbf{x})$. Using Bayes Theorem, we see

that

$$L(\mathbf{g_o}|\mathbf{x}) = \frac{L(\mathbf{x}|\mathbf{g_o})P(\mathbf{g_o})}{\sum_{GC} L(\mathbf{x}|\mathbf{g})P(\mathbf{g})}$$

where GC represents all possible genotypic configurations for that sibship type[1]. As described in Chapter 2, parental mating types and inheritance vectors combine to form the GC: for a sibship of size s there are 2^{4+2s} possible GC. The frequency of each GC, $P(\mathbf{g})$, is a simple function of allele frequency, p. For each unique sibship size s observed in the sample, the full set of GC is enumerated. Each sibship size also has a $s \times s$ residual covariance matrix, specified as above. The vectors $\mathbf{A_b}$, $\mathbf{A_w}$, $\mathbf{D_b}$ and $\mathbf{D_w}$ are calculated for each GC, both assuming parental genotypes present and absent.

Trait scores must be standardised prior to analysis, based on the population mean and variance rather than the sample mean and variance in the case of selected samples. The sibling correlation must also be specified, to allow the appropriate analysis of selected samples. Allele frequency can be specified in advance (if the population value is known) or fixed to the sample value. Alternatively, in the case of conditional analysis, it can be estimated as a free parameter.

The free parameters in the full model are a_b, a_w, d_b, d_w and p. Because the trait variance and correlation are fixed (to allow for the analysis of selected samples) it is necessary to calculate the residual components of variance, rather than estimating them as in the unconditional approach:

$$\begin{aligned} \sigma^2_S &= r - \frac{\sigma^2_A}{2} - \frac{\sigma^2_D}{4} \\ \sigma^2_N &= (1-r) - \frac{\sigma^2_A}{2} - \frac{3\sigma^2_D}{4} \end{aligned}$$

where r is the (fixed) sibling correlation and σ^2_A and σ^2_D are the variances explained

[1]In practice, the numerator is also a summed over numerous GC, although only those that are consistent with the observed sibship genotypes (and parental genotypes if available).

by the QTL, calculated from the free parameters of the model. If parents are not available, the additive and dominance QTL variances are

$$\begin{aligned}\sigma^2_A &= \frac{s+1}{2s}2pq(a_b + d_b(q-p))^2 + \frac{s-1}{2s}2pq(a_w + d_w(q-p))^2 \\ \sigma^2_D &= \frac{s+3}{4s}(2pqd_b)^2 + \frac{3s-3}{4s}(2pqd_w)^2\end{aligned}$$

where $q = 1 - p$. These coefficients are derived from elementary covariance algebra (Sham et al., 2000b): if g_i is the element of g for sibling i, the variance of the between-sibships QTL effects are

$$\begin{aligned}Var\left(\sum_{i=1}^{s}\frac{g_i}{s}\right) &= \frac{1}{s^2}\left[\sum_{i=1}^{s}Var(g_i) + \sum_{i=1}^{s}\sum_{j=i+1}^{s}2Cov(g_i,g_j)\right] \\ &= \frac{1}{s^2}\left[s(\sigma^2_A+\sigma^2_D) + s(s-1)(\frac{\sigma^2_A}{2}+\frac{\sigma^2_D}{4})\right] \\ &= \frac{s+1}{2s}\sigma^2_A + \frac{s+3}{4s}\sigma^2_D\end{aligned}$$

whereas the variance of the within QTL effect for sibling k is

$$\begin{aligned}Var\left(g_k - \sum_{i=1}^{s}\frac{g_i}{s}\right) &= Var(g_k) + Var\left(\sum_{i=1}^{s}\frac{g_i}{s}\right) - 2Cov\left(g_k, \sum_{i=1}^{2}\frac{g_i}{s}\right) \\ &= (\sigma^2_A+\sigma^2_D) + \left(\frac{s+1}{2s}\sigma^2_A + \frac{s+3}{4s}\sigma^2_D\right) \\ &\quad -2\left[\frac{\sigma^2_A+\sigma^2_D}{s} + \frac{s-1}{s}\left(\frac{\sigma^2_A}{2}+\frac{\sigma^2_D}{4}\right)\right] \\ &= \frac{s-1}{2s}\sigma^2_A + \frac{3s-3}{4s}\sigma^2_D\end{aligned}$$

If parental genotypes are available, and the between component is based on the expected sibship genotypes, this is equivalent to $s \to \infty$. For the between effects variances, $\frac{s-1}{2s}$ becomes $\frac{1}{2}$ and $\frac{s+3}{4s}$ becomes $\frac{1}{4}$. The coefficients used to weight the contributions of within effects variance are then $\frac{1}{2}$ and $\frac{3}{4}$ for additive and dominance effects, respectively. All variance components are bounded within [0,1]. The residual

covariance matrix is constructed from σ^2_S and σ^2_N, as described above.

3.3.1 Implementation

This method is implemented in the `cafe` computer program. Sibships of any size, with or without parental genotypes are read in as a pedigree file. Multi-allelic markers are analysed one allele at a time versus all other alleles. A model is specified for both the alternate (`--alt`) and null (`--null`) hypotheses as follows. The letters `b`, `w` and `f` indicate between, within and fixed (i.e. between = within) components are to be estimated. Specifying `0` means that no association components are estimated. The robust test is therefore `--alt bw --null b` (or alternatively, `--alt w -null 0`). Using uppercase letters indicates dominance, rather than additive, effects.

3.4 Sample selection

As for the QTL linkage test, discussed in Chapter 2, an index of potential informativeness for each sibship is the sum over all GC, $\sum_i P_i t_i$ where P_i is the probability of observing that GC given the trait scores and the assumed true model and t_i is the contribution to the test statistic that would be obtained if that GC were true. The probability of observing a particular GC i conditional on trait scores $\mathbf{x}$ can be calculated using Bayes Theorem as

$$P_i = (GC_i|\mathbf{x}) = \frac{L(\mathbf{x}|GC_i)P(GC_i)}{\sum_j L(\mathbf{x}|GC_j)P(GC_j)}$$

where $L(\mathbf{x}|GC_j)$ is calculated as shown above, under the true model.

The potential informativeness of a phenotyped sibship for a number of different tests within the Fulker association model can be calculated, by appropriately specifying the expected means vector and residual covariance matrix for each GC under an assumed true model. If an effect is not modelled in the means vector, it will have an

Model	μ	Σ	Interpretation
0	0	0	No QTL effect
N . .	0	BW	No association: all QTL effects modelled in covariance
BW	BW	0	Total association
B	B	W	Between-sibships association
W	W	B	Within-sibships association

Table 3.3: Possible models within the Fulker association framework.

impact on the residual covariance matrix. Table 3.3 shows five different scenarios: **0**, no QTL effect; **N**, QTL effect but not modelled in means; **BW**, QTL effect modelled in means; **B**, only between QTL effect modelled in means; **W**, only within QTL effect modelled in means.

The means vector μ is constructed as above, including only the between, only within, both or neither components of association as necessary. If a QTL effect is present but not modelled in the means, it will have an impact on the covariance structure, which becomes

$$[\Sigma_{\mathbf{BW}}]_{ij} = \begin{cases} \sigma_A^2 + \sigma_D^2 + \sigma_S^2 + \sigma_N^2 & \text{for } i = j \\ \frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4} + \sigma_S^2 & \text{for } i \neq j \end{cases}$$

If the QTL effect is modelled in the means (or not present), then the expected covariance matrix will be simply

$$[\Sigma_{\mathbf{0}}]_{ij} = \begin{cases} \sigma_S^2 + \sigma_N^2 & \text{for } i = j \\ \sigma_S^2 & \text{for } i \neq j \end{cases}$$

If only the within component is modelled in the means, the between component will impact the expected values of the covariance matrix

$$[\Sigma_{\mathbf{B}}]_{ij} = \begin{cases} \frac{s+1}{2s}\sigma_A^2 + \frac{s+3}{4s}\sigma_D^2 + \sigma_S^2 + \sigma_N^2 & \text{for } i = j \\ \frac{s+1}{2s}\sigma_A^2 + \frac{s+3}{4s}\sigma_D^2 + \sigma_S^2 & \text{for } i \neq j \end{cases}$$

whereas if only the between component is modelled in the means, the expected within covariance matrix is

$$[\Sigma_{\mathbf{W}}]_{ij} = \begin{cases} \frac{s-1}{2s}\sigma_A^2 + \frac{3s-3}{4s}\sigma_D^2 + \sigma_S^2 + \sigma_N^2 & \text{for } i = j \\ \frac{-1}{2s}\sigma_A^2 + \frac{-3}{4s}\sigma_D^2 + \sigma_S^2 & \text{for } i \neq j \end{cases}$$

These coefficients for σ_A^2 and σ_D^2 were derived above, with the exception of the off-diagonal element for pair (k, l) in $\Sigma_{\mathbf{W}}$,

$$\begin{aligned} Cov\left[\left(g_k - \sum_{i=1}^{s}\frac{g_i}{S}\right), \left(g_l - \sum_{i=1}^{s}\frac{g_i}{S}\right)\right] &= Cov(g_k, g_l) + Var\left(\sum_{i=1}^{s}\frac{g_i}{s}\right) \\ &\quad -Cov\left(g_k, \sum_{i=1}^{s}\frac{g_i}{s}\right) - Cov\left(g_l, \sum_{i=1}^{s}\frac{g_i}{s}\right) \\ &= \left(\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4}\right) + \left(\frac{s+1}{2s}\sigma_A^2 + \frac{s+3}{4s}\sigma_D^2\right) \\ &\quad -2\left[\frac{\sigma_A^2 + \sigma_D^2}{s} + \frac{s-1}{s}\left(\frac{\sigma_A^2}{2} + \frac{\sigma_D^2}{4}\right)\right] \\ &= \frac{-1}{2s}\sigma_A^2 + \frac{-3}{4s}\sigma_D^2 \end{aligned}$$

As outlined in Table 3.3, several models can be specified: total association

$$\ln L_{BW} = -\frac{1}{2}[\ln|\Sigma_0| + (\mathbf{x} - \mu_{\mathbf{BW}})'\Sigma_0^{-1}(\mathbf{x} - \mu_{\mathbf{BW}})]$$

between sibship association

$$\ln L_B = -\frac{1}{2}[\ln|\Sigma_{\mathbf{W}}| + (\mathbf{x} - \mu_{\mathbf{B}})'\Sigma_{\mathbf{W}}^{-1}(\mathbf{x} - \mu_{\mathbf{B}})]$$

within sibship association

$$\ln L_W = -\frac{1}{2}[\ln|\Sigma_{\mathbf{B}}| + (\mathbf{x} - \mu_{\mathbf{W}})'\Sigma_{\mathbf{B}}^{-1}(\mathbf{x} - \mu_{\mathbf{W}})]$$

and no QTL effect

$$\ln L_N = -\frac{1}{2}[\ln|\boldsymbol{\Sigma}_\mathbf{N}| + (\mathbf{x}-0)'\boldsymbol{\Sigma}_\mathbf{N}^{-1}(\mathbf{x}-0)].$$

Based on the unconditional association test, the expected informativeness for a sibship for the total association test is given by

$$\sum_{GC}(2\ln L_{BW} - 2\ln L_N)P(GC|\mathbf{x})$$

whereas for the conditional test it is

$$\sum_{GC} 2\left(\ln\frac{L_{BW}P(GC)}{\sum_{GC}L_{BW}P(GC)} - \ln\frac{L_NP(GC)}{\sum_{GC}L_NP(GC)}\right)P(GC|\mathbf{x})$$

Similarly, for the first specification of the robust within test (W1), the index is

$$\sum_{GC} 2\left(\ln\frac{L_{BW}P(GC)}{\sum_{GC}L_{BW}P(GC)} - \ln\frac{L_BP(GC)}{\sum_{GC}L_BP(GC)}\right)P(GC|\mathbf{x})$$

whereas the second specification (W2) is

$$\sum_{GC} 2\left(\ln\frac{L_WP(GC)}{\sum_{GC}L_WP(GC)} - \ln\frac{L_NP(GC)}{\sum_{GC}L_NP(GC)}\right)P(GC|\mathbf{x})$$

These procedures are implemented in the SEA computer program.

3.4.1 Non-independence of sibships for association informativeness

Broadly speaking, selection for linkage enriches the sample for sibling pairs that are IBD 0 or IBD 2 at the QTL. Significantly, this strategy would still work if the pairs selected were either all IBD 0 or all IBD 2, because the test for linkage implicitly

fixes the population mean IBD value, i.e. at 1. Similarly, selection for association will enrich the sample for individuals homozygous for trait loci, say 1/1 or 2/2. However, the strategy would not work if all individuals selected were 1/1 or all were 2/2. That is, the informativeness of a sibship for association is not independent of the genotypes (and therefore trait scores) of all other sibships in the sample. As an extreme example, if during selection a sample contains only 1/1 individuals, then adding another 1/1 individual will not increase power, whilst an 1/2 individual would. Applying the conditional test, however, it is possible to fix the allele frequency to the known population value. In this case, it would actually be possible for a selected sample with no genetic variation to show a positive association.

3.4.2 Properties of selection for association

Unselected samples

Before considering the properties of selected samples, it is worth briefly looking at the properties of unselected samples, using both SEA and GPC, which give the expected test statistic for unselected sibships under the Fulker variance components association model via an analytic approach.

Using SEA, the expected sample noncentrality parameter (NCP) for a test of association can be calculated as follows: 1) simulate a large dataset (e.g. 50,000 sibships) under the null of no QTL effect but multivariate normal with the desired sibling correlation, 2) calculate the expected contribution to the NCP for each sibship for a particular assumed true model, 3) sum the expected contributions for each test over all sibships, 4) scale by desired sample size (e.g. by 100/50,000 for a sample size of 100 sibships).

Several conditions are considered: 1200 singletons, 600 pairs, 400 trios and 300 quads; three different sibling correlations, $r = 0.2$, 0.5 and 0.8; QTL accounting for 2 and 10% of the trait variance. The test locus is a diallelic QTL with equifrequent

alleles. In all cases, the expected NCP for the between association test, the robust within association test and the total test is determined. As between and within components of association are, in unselected samples, orthogonal, they sum to give the total association informativeness.

Several expected observations were confirmed (c.f. Sham et al., 2000b). Firstly, the results based on GPC and SEA match very closely in all cases, confirming the correct implementation of SEA. Secondly, the expected NCP is roughly linearly related to the proportion of variance explained by the QTL: the test statistics for a 10% QTL are typically 5 times greater than those for a 2% QTL. For singletons, the within component NCP is necessarily 0, but this increases with increasing sibship size and increasing sibling correlation. In contrast, the between component decreases with increasing sibship size and increasing sibling correlation. For the total association test, if the sibling correlation $r < 0.5$ then smaller sibships are favoured; if $r > 0.5$ then larger sibships are favoured.

Profiles of selected pairs

For the three different sibling correlations (0.2, 0.5 and 0.8), Figure 3.2 illustrates the impact of selection on sibling pairs. The top row of plots represent the unselected samples; the second row of plots represent the 5% of the sample most informative for total association; the third row represent the 5% most informative for between association; the fourth row represent the 5% most informative for within association. Clearly, pairs informative for the between component are extreme concordant high and low pairs; pairs informative for the within component have discordant trait scores. Whether or not the pairs informative for the total association show a preponderance of concordant or discordant pairs depends on the sibling correlation, as mentioned above. The index of informativeness for total association appears to be similar to the Mahalanobis distance.

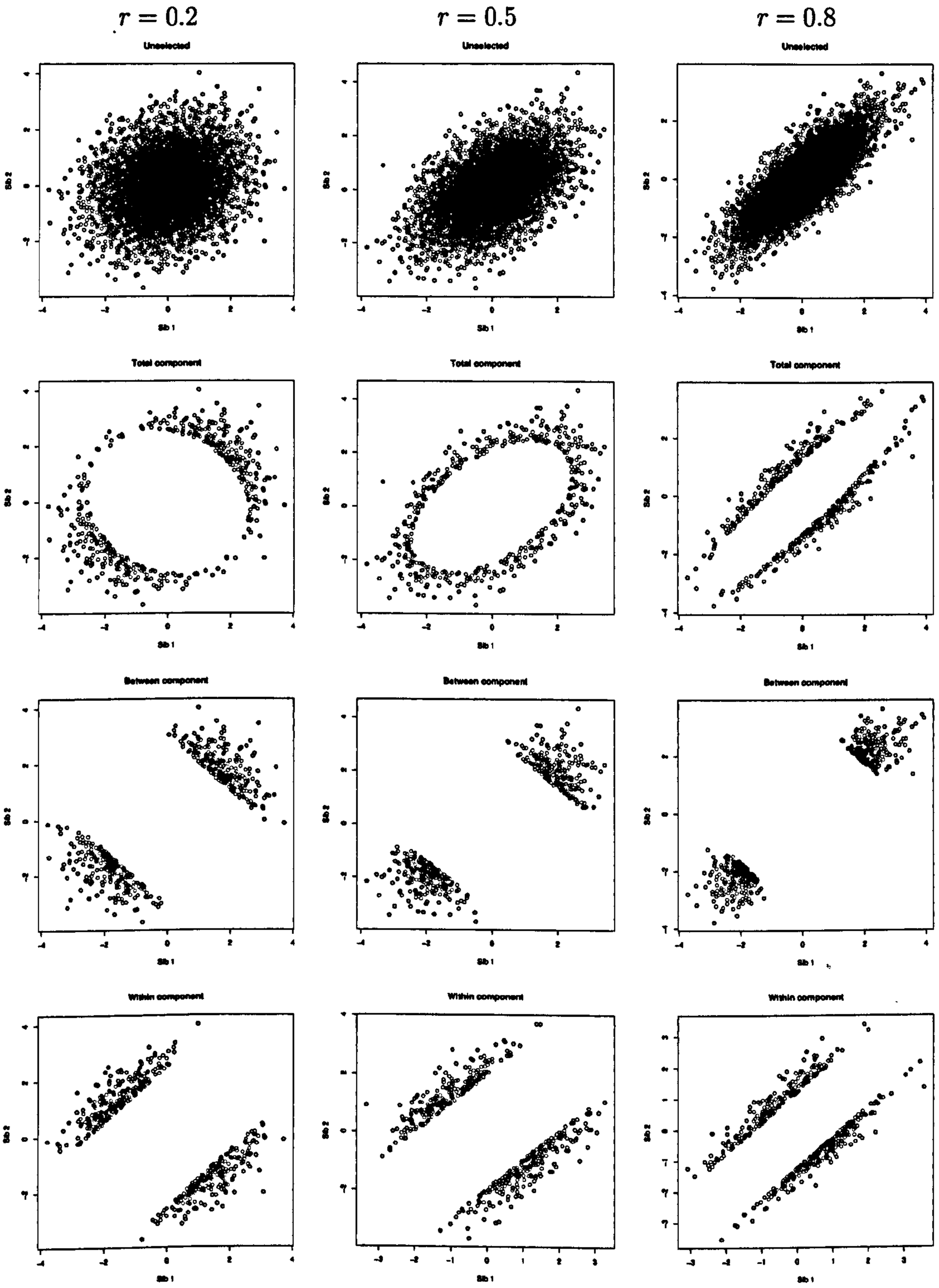


Figure 3.2: Profiles of sibling pairs selected for association.

Efficiency of selected samples

For a 2% QTL, Figure 3.3 displays the different NCPs for both unselected and selected samples for different sibship sizes and different sibling correlations. The left-hand column of plots illustrates more clearly the observations made in the previous section regarding the relative merit of the between and within components as a function of sibship size and sibling correlation.

The efficiency of selection is illustrated in Table 3.4 (results only given for a 2% QTL; similar results obtained for a 10% QTL). For a 5% selection, the Table gives the proportion of the unselected sample NCP retained, for the three association tests (B, W and BW) for samples selected on the basis of B, W or BW informativeness (i.e. 9 conditions). For samples selected on the basis on total (BW) association, between 27% (singletons) and 15% (quads) of the BW NCP is retained. Therefore the efficiency of selection for association is, as expected, much less than for linkage. The figure for the larger sibships is probably unrealistically low, as sibships are selected whole in this procedure. If uninformative individuals were to be 'dropped' from otherwise informative sibships, then the efficiency for selection should not decrease with sibship size in the same way. For sibships selected on BW, the B signal is better retained in sibships when the sibling correlation is low (and obviously in singletons); conversely, for sibships selected on BW, the W signal is better retained in sibships when the sibling correlation is high.

For sibships selected on B, the amount of signal retained on B is uniformly high (27%) across all sibship sizes and sibling correlations. However, the W signal is uniformly low (5%) for sibships selected on B, and the BW signal decreases with increasing sibship size and sibling correlation. In contrast, for sibships selected on W, although the signal retained for B is uniformly low (5%), the signal for BW increases within increasing sibship size and sibling correlation. The signal for W decreases with increasing sibship size, however (although, again, this is not defined for singletons).

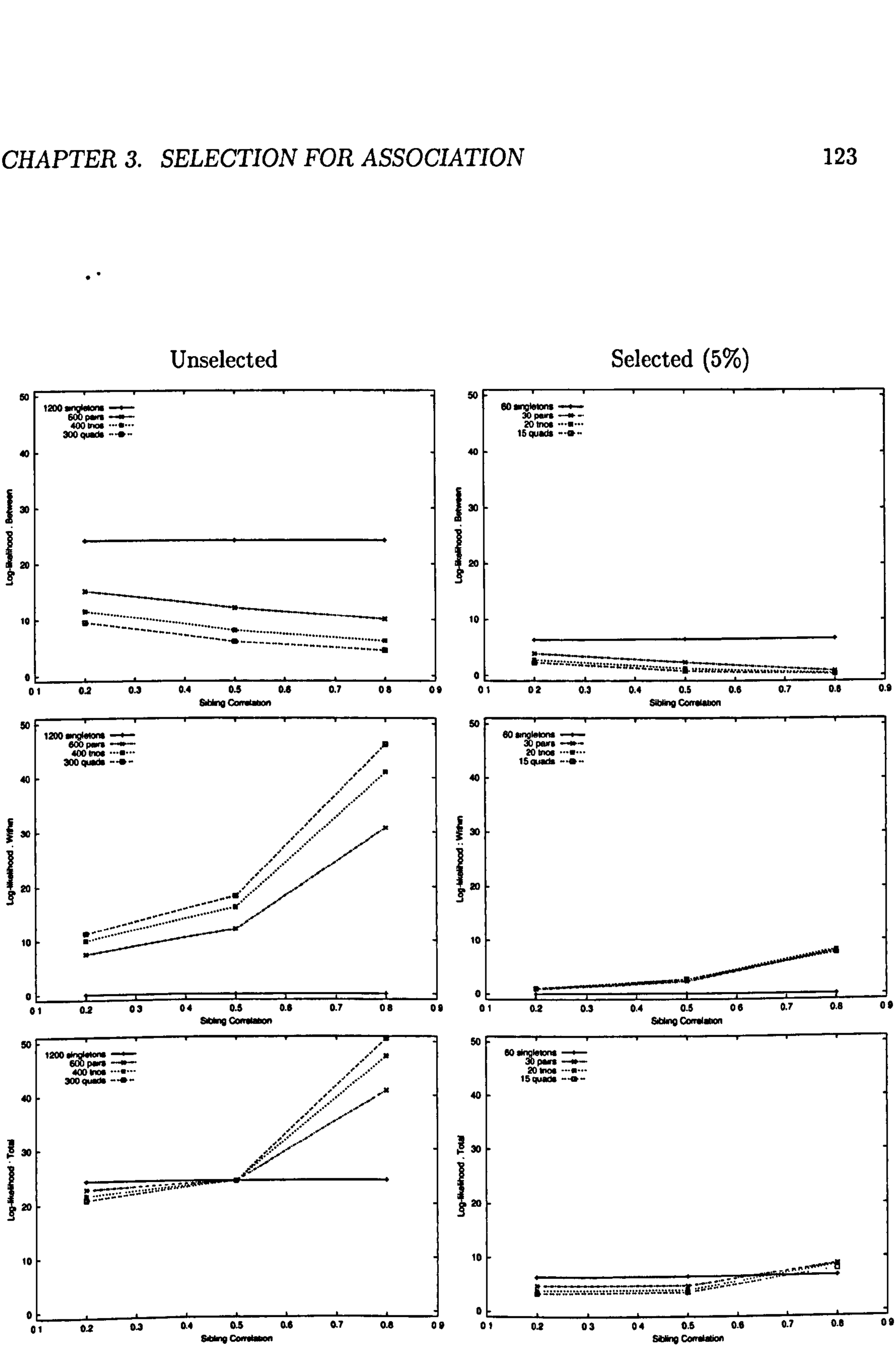


Figure 3.3: Unselected and selected sample NCPs for Fulker association model; 2% QTL.

		Sel on BW			Sel on B			Sel on W		
	r	B	W	BW	B	W	BW	B	W	BW
2% QTL										
60 singletons	0.2	0.27	0.00	0.27	0.27	0.00	0.27			
	0.5	0.27	0.00	0.27	0.27	0.00	0.27			
	0.8	0.28	0.00	0.28	0.28	0.00	0.28			
30 sib pairs	0.2	0.26	0.12	0.21	0.28	0.05	0.20	0.05	0.28	0.12
	0.5	0.19	0.20	0.19	0.27	0.04	0.16	0.05	0.28	0.16
	0.8	0.09	0.27	0.22	0.27	0.04	0.10	0.04	0.27	0.22
20 sib trios	0.2	0.25	0.10	0.18	0.27	0.04	0.17	0.05	0.20	0.12
	0.5	0.16	0.17	0.17	0.27	0.05	0.12	0.05	0.20	0.15
	0.8	0.07	0.20	0.18	0.28	0.05	0.08	0.05	0.20	0.18
15 sib quads	0.2	0.24	0.09	0.16	0.27	0.05	0.15	0.05	0.17	0.11
	0.5	0.14	0.15	0.15	0.28	0.05	0.11	0.04	0.17	0.14
	0.8	0.06	0.17	0.16	0.28	0.05	0.07	0.05	0.17	0.16

Table 3.4: Efficiency of selection for association: for a 2% QTL, the figures represent the proportion of total NCP retained from 5% selection.

In general, given that the within component of association is robust to stratification effects, it would seem desirable to select on W. However, for the selection of singletons without parents, clearly one must select on B(=BW). If the sibling correlation is high, then selecting on BW is virtually identical to selecting on W in any case. Although selection for association is not as efficient as for linkage this only reflects the staggering inefficiency of linkage in unselected samples. As with linkage, QTL variance has no real effect on selection, although it will, of course, have a massive effect on power.

Unequal allele frequency and dominance

All scenarios so far have assumed equal allele frequencies for the additive QTL. This section assesses the impact of assuming the 'base model' of equal allele frequencies and no dominance during selection when the true model is different.

The revised profiles of the most informative 5% of sibling pairs are displayed in Figure 3.4. Three different QTL scenarios correspond to the three columns of plots; the three rows correspond to selection on BW, B and W moving from top to bottom. The increaser allele has frequency p, so that in the first case (left-most column of plots) the allele associated with higher trait scores is rare (10%). As seen in selection for linkage, the end of the distribution associated with the less common allele will

		Selection (BW) model	
True model	Test	True	Base
Additive, $p = 0.5$	B+W	18	18
(Base)	B	17	17
	W	20	20
Additive, $p = 0.1$	B+W	26	22
	B	25	20
	W	25	24
Dominance, $p = 0.1$	B+W	24	21
	B	21	18
	W	24	23
Dominance, $p = 0.9$	B+W	56	55
	B	62	71
	W	68	62

Table 3.5: Impact of model misspecification: the NCP expressed as a percentage of total NCP in the overall sample.

be more informative: more variation at this end can be explained in terms of the genotype. That is, at the common end, all individuals will be homozygous for the common allele. This is most clearly seen in selection on B and BW.

If the rare allele also has dominant gene action, the pattern of selection remains unchanged (middle column), indicating that an additive approximation still works well. However, if the rare allele is recessive (rightmost column, i.e. 'common dominant') a distinctly different pattern of results is obtained. Whether selecting for BW, B or W association, in all cases sibling pairs with concordant low scores are preferentially selected over sibling pairs with concordant high scores, as there is now a very strong tendency to oversample the highly informative (but very rare) recessive homozygotes.

Table 3.5 shows the percentage of the NCP retained in a 5% selected sample, depending on whether selection occurred under the base model or the true model. In all cases, these figures are based on a sample of pairs with a sibling correlation of 0.50 and a 10% QTL. Selection was always based on the total (BW) index.

As seen in the previous Table, approximately 18% of the total information for association is retained when 5% of sibling pairs with a sibling correlation of 0.5 are

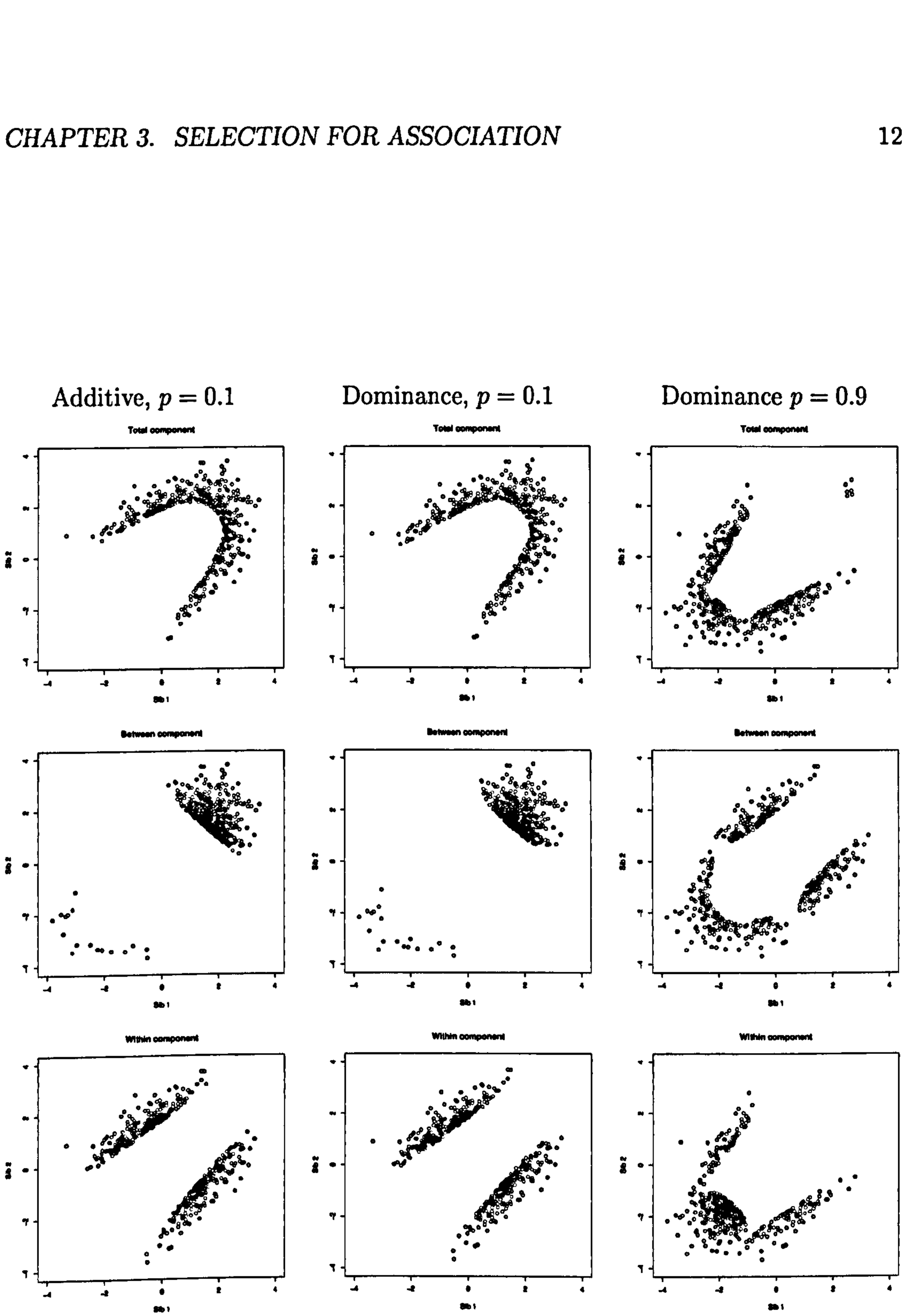


Figure 3.4: Impact of unequal allele frequencies and dominance.

selected. Under the next two models, an even higher proportion of information is retained, even if selection incorrectly assumes the base model.

The fourth (rare recessive) condition has much smaller absolute NCP. That is, even though the proportion of variance accounted for is the same (10%), the approximate power equations break down in this case. The proportion of information retained is even higher in these situations[2]. It makes intuitive sense that the less powerful the unselected sample, the more efficient selection will be, i.e. as the original lack of power is due to the majority of individuals not contributing. In general, these results support the use of the base model in selection.

Informativeness for association in samples selected for linkage.

Finally, it is of interest to examine how efficient a linkage selection strategy is for association. A common genome-scan study design would involve screening with a first phase of linkage followed by association for the linkage 'hotspots'. In this case, the samples will tend to be initially selected on the basis of the informativeness for linkage. How informative will these samples be for association?

In all cases the base model was assumed along with a 10% QTL. Unselected samples of sibling pairs, trios and quads (similar sample sizes to previous) were generated from which the most informative 5% were selected. Selection occurred either on the basis of informativeness for linkage, total association, between association or within association. Figure 3.5 plots the total expected NCP for these selected samples: the four columns of plots represent these four selection strategies; the three rows of plots represent pairs, trios and quads.

It is immediately clear that selecting for linkage and selecting for within association have very similar profiles, especially for larger sibships. These profiles are, in turn,

[2]The base model B is actually greater than the true model B efficiency in this condition (71 vs 62) – this is only because true model selection was based on BW, not B; therefore B and W figures may not be optimal.

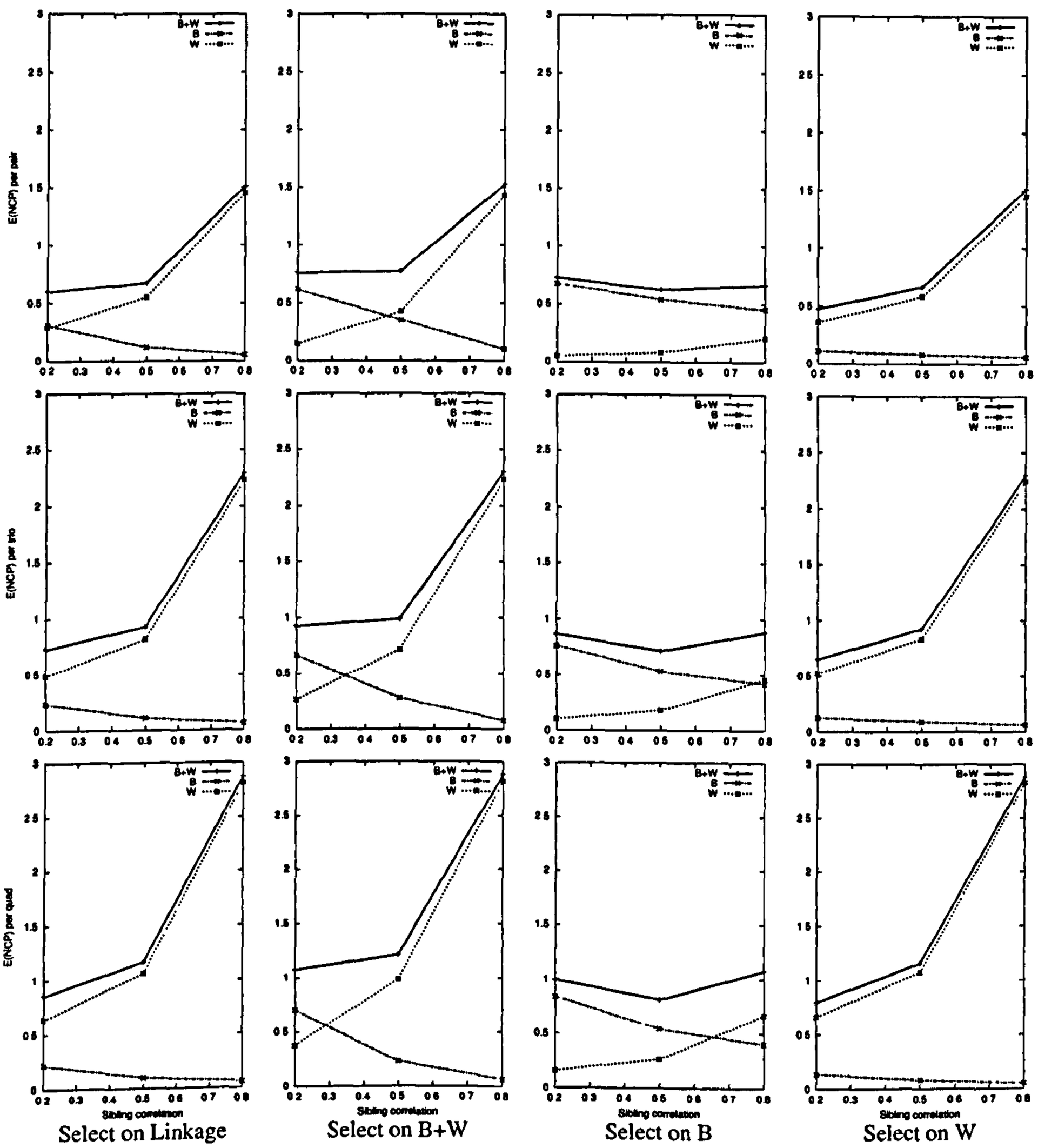


Figure 3.5: Selection schemes for linkage and association.

similar to the total association profile, at least for BW and W at reasonably high sibling correlations. So, the sibships that are selected for linkage will also be close to optimal for the subsequent association analyses, especially when the robust within sibship test is used. The use of larger sibships is clearly advantageous in both contexts.

3.5 Simulation study of QTL association in selected samples

3.5.1 Overview of simulations

The properties of the conditional approach to family-based association analysis are investigated in a series of simulations. Samples of singletons and sibling pairs (with and without parental genotypes) are generated with the test locus either having no effect on the trait (the null) or accounting for 2% of the phenotypic variance in an additive manner (the alternate). In all cases, the test locus is a diallelic marker with equal allele frequencies. The singleton samples consist of 1200 individuals; the pair samples consist of 600 pairs.

For singletons and pairs, two simple selection schemes are applied to the data, illustrated in Figure 3.6. The top row illustrates the full sample case for singletons (F), an "extreme high / extreme low" scheme (A) and an "extreme low / random controls" scheme (B). Sampling schemes A and B select 5% of the unselected sample (i.e. 60 individuals). The bottom row of plots indicates the equivalent procedures for sibling pairs, with different selection schemes applied: discordant pairs (A) and concordant high pairs (B). Again, 5% of pairs are selected under A and B (30 pairs).

In all cases, the unselected samples have a mean of 0 and a variance of 1; in the case of sibling pairs, the sibling correlation is $r = 0.5$ (simulations were also conducted with $r = 0.2$ and $r = 0.8$, giving largely similar patterns of results with respect to

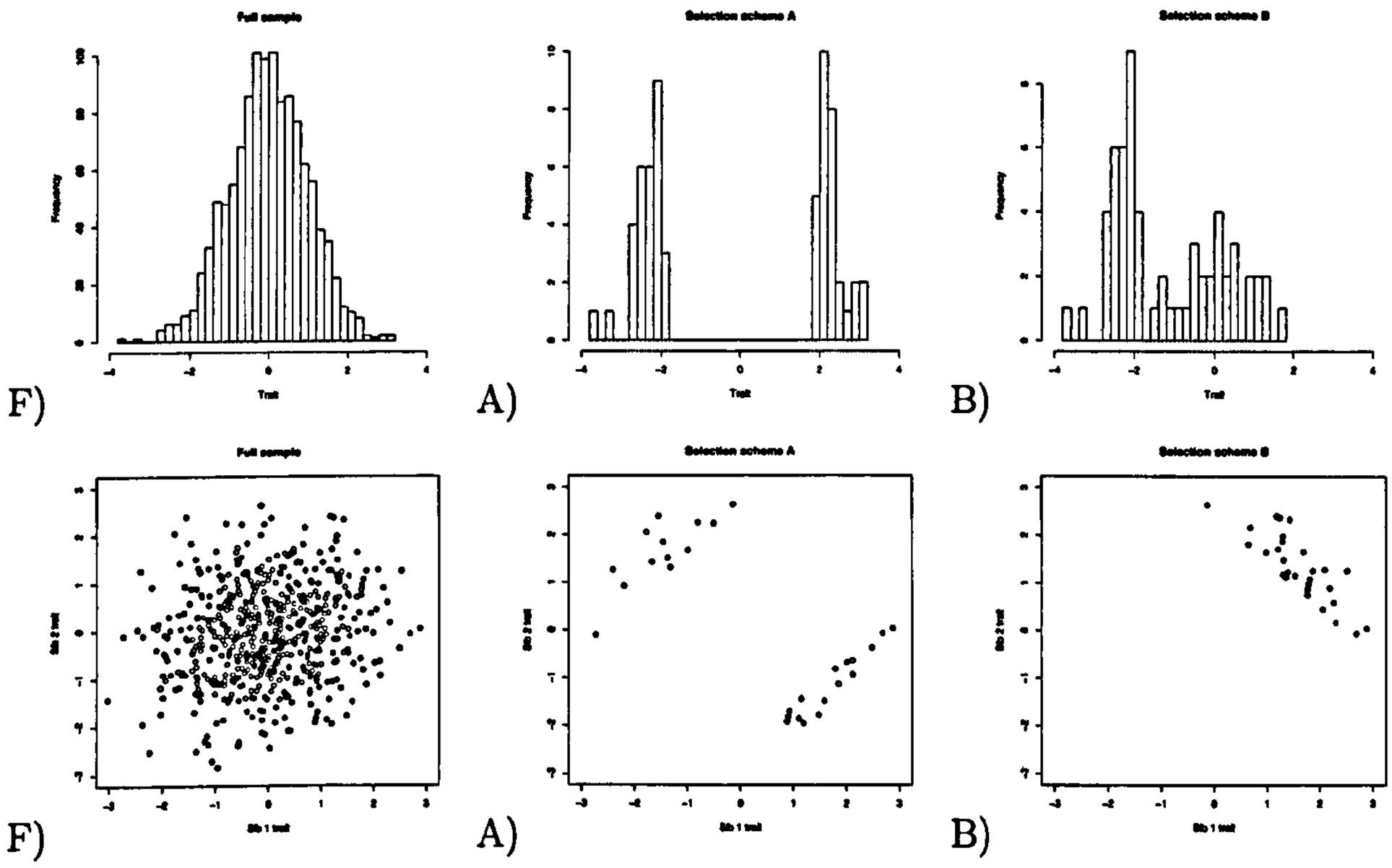


Figure 3.6: Selection schemes applied to singleton samples (top row) and sibling pair samples (bottom row).

the performance of different analytic approaches). In conditions F, A and B, the mean, variance and sibling correlation are fixed to these values, for both selected and unselected analyses. A further condition emulates the effect of mis-specifying these parameters: in the M condition (full sample only) the trait scores are multiplied by 2 prior to analysis. Therefore, the trait variance is increased 4-fold, although it is still fixed to 1 in analysis, meaning that observations will appear to be more extreme than they actually are.

The data are analysed under a number of different conditions. The use of a conditional approach versus the standard approach is varied (Cond). Three different tests (Test) are employed also: a total associated test (T) and two within sibship tests (W1 and W2). The T test equates free parameters a_B and a_W under the alternate and fixes them both to 0 under the null. The W1 test estimates both a_B and a_W separately under the alternate but only a_B under the null. The W2 test estimates only a_W under the alternate and fixes it to 0 under the null. Additionally, whether

or not parental genotypes were included in the analysis (Par) is varied.

The treatment of allele frequency is also varied in the simulations. For the main results given below, the allele frequency parameter is fixed to its population value (i.e. its expected value in the unselected sample). The two other conditions are to fix the allele frequency to the sample value (whether the sample is selected or not) or to treat allele frequency as a free parameter to be estimated jointly with the other QTL parameters. This latter option is only possible when a conditional approach is adopted (i.e. as then the likelihood of observing genotype conditional on trait is a function of allele frequency). For the majority of simulation results reported here, fixing allele frequency to the population value is chosen for convenience. The effect of fixing allele frequency to the sample value or estimating it is explored in a separate section below.

For each of the four conditions (singletons versus pairs, no QTL versus QTL) 1000 samples with parents and a separate 1000 samples without parents are generated, each analysed under a total of 78 conditions (not all shown in the results).

3.5.2 Robustness under the null

Table 3.6 gives the results for the full sample with data simulated under the null hypothesis of no QTL effect. The first four columns Sel, Cond, Test and Par indicate the type of sample and analytic method used (selection scheme, conditioning or not, test statistic and parental genotypes). For singletons ($s = 1$) and pairs ($s = 2$) the likelihood ratio test statistic (LRT), Type I error rate and estimated total QTL variance are presented. The expected LRT under the null is 1 (all tests are 1 degree of freedom tests). The expected Type I error rate is 5%, as $\alpha = 0.05$. Results are highlighted in **bold** type if the LRT grossly departs from the expected value (difference greater than twice the standard error which is $2\sqrt{2/1000} \approx 0.1$).

For the full sample analyses (Sel = F) the results are close to their expected

				LRT		Type I		VC	
Sel	Cond	Test	Par	$s=1$	$s=2$	$s=1$	$s=2$	$s=1$	$s=2$
F	N	W1	N		1.03		5.4%		0.2%
F	N	W2	N		1.03		5.4%		0.1%
F	N	T	N	1.01	1.04	5.7%	5.6%	0.1%	0.1%
F	N	W1	Y	1.00	0.96	4.5%	5.2%	0.2%	0.2%
F	N	W2	Y	1.00	0.96	4.2%	5.2%	0.2%	0.1%
F	N	T	Y	0.99	0.95	4.9%	4.5%	0.1%	0.1%
F	Y	W1	N		1.04		5.7%		0.2%
F	Y	W2	N		1.04		5.7%		0.1%
F	Y	T	N	1.01	1.04	5.7%	5.6%	0.1%	0.1%
F	Y	W1	Y	1.00	0.96	4.5%	5.2%	0.2%	0.2%
F	Y	W2	Y	1.00	0.96	4.2%	5.2%	0.2%	0.1%
F	Y	T	Y	0.99	0.95	4.9%	4.5%	0.1%	0.1%
M	N	W1	N		**1.65**		**12.6%**		0.0%
M	N	W2	N		**1.65**		**12.6%**		0.0%
M	N	T	N	1.01	1.04	5.4%	5.5%	0.0%	0.0%
M	N	W1	Y	1.00	**1.18**	4.3%	**7.9%**	0.0%	0.0%
M	N	W2	Y	1.00	**1.18**	4.3%	**7.9%**	0.0%	0.0%
M	N	T	Y	0.99	0.95	4.9%	4.7%	0.0%	0.0%
M	Y	W1	N		1.04		5.8%		0.0%
M	Y	W2	N		1.04		5.8%		0.0%
M	Y	T	N	1.01	1.04	5.5%	5.5%	0.0%	0.0%
M	Y	W1	Y	1.00	0.96	4.3%	5.0%	0.0%	0.0%
M	Y	W2	Y	1.00	0.96	4.3%	4.9%	0.0%	0.0%
M	Y	T	Y	0.98	0.95	4.9%	4.7%	0.0%	0.0%

Table 3.6: Simulation study results: full samples under the null.

values. The most striking results in Table 3.6 reflect the effects of mis-specifying the trait variance (Sel = M). In this case, all within association tests in pairs are anti-conservative if the standard unconditional approach is adopted (shown in **bold**). In these cases, the Type I error rate can more than double. Under the conditional approach, the results are much closer to their expected values.

Table 3.7 presents the results for selected samples, schemes A and B, under the null. Again, striking departures from the expected results are highlighted in **bold**. Immediately clear is a pattern of results suggesting that the use of within sibship association tests in selected samples can be lead to highly liberal test statistics if a

				LRT		Type I / power		VC	
Sel	Cond	Test	Par	$s=1$	$s=2$	$s=1$	$s=2$	$s=1$	$s=2$
A	N	W1	N		**2.86**		**26.7%**		1.3%
A	N	W2	N		**2.85**		**26.6%**		2.2%
A	N	T	N	1.03	1.08	5.9%	4.5%	0.3%	0.6%
A	N	W1	Y	1.00	**1.62**	5.1%	**11.6%**	0.6%	0.9%
A	N	W2	Y	1.00	**1.62**	4.9%	**11.7%**	0.6%	1.5%
A	N	T	Y	0.96	1.01	4.7%	4.9%	0.3%	0.5%
A	Y	W1	N		1.10		5.7%		3.0%
A	Y	W2	N		1.10		5.7%		0.4%
A	Y	T	N	1.03	1.08	5.9%	4.5%	0.3%	0.6%
A	Y	W1	Y	1.01	0.99	5.1%	5.0%	0.6%	3.2%
A	Y	W2	Y	1.00	0.99	4.9%	4.7%	0.6%	0.6%
A	Y	T	Y	0.96	1.01	4.7%	4.9%	0.3%	0.5%
B	N	W1	N		**0.55**		**1.0%**		1.2%
B	N	W2	N		**0.54**		**1.0%**		0.4%
B	N	T	N	1.02	1.07	4.5%	5.4%	0.5%	0.7%
B	N	W1	Y	0.98	**0.84**	4.1%	**3.3%**	1.1%	1.4%
B	N	W2	Y	0.97	**0.84**	4.1%	**3.2%**	1.0%	0.9%
B	N	T	Y	0.98	1.02	4.6%	5.4%	0.5%	0.6%
B	Y	W1	N		1.09		6.1%		1.8%
B	Y	W2	N		1.10		6.1%		2.4%
B	Y	T	N	1.02	1.07	4.5%	5.4%	0.5%	0.7%
B	Y	W1	Y	0.98	1.04	4.1%	5.3%	1.1%	1.3%
B	Y	W2	Y	0.97	1.03	4.1%	5.1%	1.0%	1.4%
B	Y	T	Y	0.98	1.02	4.6%	5.4%	0.5%	0.6%

Table 3.7: Simulation study results: selected samples under the null.

standard, unconditional approach to analysis is adopted. For selection scheme A, the Type I error rate is over 25% (the within sibship test for sibling pairs without parental genotypes). In contrast, results when conditioning are much closer to their expected values. For selection scheme A, the results for pairs without parents when conditioning show slightly inflated average LRTs, although the observed Type I error rates do not seem to deviate from 5%. For selection scheme B, a similar pattern emerges, except that in this case the unconditional approach produces conservative results for the within test.

The associated estimated variance explained by the within tests is often too high,

particularly for Sel = A. Note that the variance explained is derived from the alternate model: in the case of the W1 method this contains both between and within components. Indeed, it is well known that the between and within components are no longer necessarily orthogonal in selected samples (Abecasis et al., 2001a) and that the W2 test should be preferred in this case.

3.5.3 Power under the alternate

Tables 3.8 and 3.9 report simulation results when a QTL explains 2% of the total phenotypic variance. Conditions that were highlighted **bold** under the null to reflect a liberal test are still highlighted in the Tables under the alternate. Table 3.8 shows the results for full samples and when the sample variance has been misspecified. In all cases, power is high. Importantly, the W1 and W2 tests have similar power whether or not parental genotypes are unavailable.

A further important result is that the power of the conditional test seems unaffected by the misspecification of the variance (Sel = M) – that is, the LRT values are very similar to the correctly specified full sample case. In other words, conditioning protects against false positives in misspecified samples, but still retains full power under the alternate hypothesis. This would indicate that the exact value the variance is fixed to is not critical. Of course, the estimated variance explained is biased in this case however (i.e. $0.02/4 = 0.005$ in this case). Further simulations are required to examine the effect of mis-specifying the mean and/or correlation.

Table 3.9 shows the results for selected samples with a 2% QTL. For the total association test T, results are similar whether or not one conditions on trait values. Reasonable power is maintained when conditioning is applied to the within tests.

				LRT		Power		VC	
Sel	Cond	Test	Par	$s=1$	$s=2$	$s=1$	$s=2$	$s=1$	$s=2$
F	N	W1	N		13.34		94.7%		2.1%
F	N	W2	N		13.16		94.4%		1.0%
F	N	T	N	24.72	25.26	99.6%	100.0%	2.0%	2.1%
F	N	W1	Y	12.67	16.83	93.1%	98.1%	2.1%	2.1%
F	N	W2	Y	12.54	16.69	93.1%	97.9%	2.1%	2.1%
F	N	T	Y	24.58	24.79	99.8%	99.7%	2.0%	2.0%
F	Y	W1	N		13.23		94.6%		2.2%
F	Y	W2	N		13.24		94.6%		1.2%
F	Y	T	N	24.72	25.26	99.6%	100.0%	2.0%	2.1%
F	Y	W1	Y	12.67	16.78	93.1%	97.9%	2.1%	2.2%
F	Y	W2	Y	12.54	16.72	93.1%	97.9%	2.1%	2.1%
F	Y	T	Y	24.58	24.79	99.8%	99.7%	2.0%	2.0%
M	N	W1	N		**21.12**		**98.1%**		0.6%
M	N	W2	N		**21.05**		**98.1%**		0.8%
M	N	T	N	24.54	25.09	99.6%	100.0%	0.5%	0.5%
M	N	W1	Y	12.54	**20.54**	92.7%	**98.7%**	0.5%	0.5%
M	N	W2	Y	12.49	**20.49**	92.8%	**98.7%**	0.5%	0.8%
M	N	T	Y	24.41	24.64	99.8%	99.7%	0.5%	0.5%
M	Y	W1	N		13.18		94.2%		0.5%
M	Y	W2	N		13.14		94.2%		0.3%
M	Y	T	N	24.63	25.16	99.6%	100.0%	0.5%	0.5%
M	Y	W1	Y	12.61	16.72	92.9%	97.9%	0.5%	0.5%
M	Y	W2	Y	12.50	16.64	92.8%	97.9%	0.5%	0.5%
M	Y	T	Y	24.49	24.70	99.8%	99.7%	0.5%	0.5%

Table 3.8: Simulation study results: full samples under the alternate.

3.5.4 Estimation of allele frequencies

All of the simulations reported above were also conducted with allele frequency fixed to the sample value, and with allele frequency as a free parameter (under both the alternate and null hypotheses), i.e. as well as being fixed to the unselected population value. For the conditions simulated, i.e. an additive 2% QTL with equal allele frequencies, the results did not differ substantially by treatment of allele frequency. However, as the QTL effect size increases, then selection can distort the sample allele frequency estimates. This section considers the special cases where selection means that the sample allele frequency deviates from the population allele frequency.

				LRT		Power		VC	
Sel	Cond	Test	Par	$s=1$	$s=2$	$s=1$	$s=2$	$s=1$	$s=2$
A	N	W1	N		**11.21**		**74.5%**		4.4%
A	N	W2	N		**11.17**		**74.4%**		8.4%
A	N	T	N	7.40	4.95	73.3%	52.4%	2.3%	2.5%
A	N	W1	Y	4.17	**7.29**	42.7%	**63.3%**	2.5%	3.5%
A	N	W2	Y	4.13	**7.27**	42.4%	**63.3%**	2.5%	6.6%
A	N	T	Y	7.35	4.86	73.5%	50.5%	2.2%	2.5%
A	Y	W1	N		4.44		47.4%		4.8%
A	Y	W2	N		4.45		47.5%		1.4%
A	Y	T	N	7.42	4.95	73.3%	52.4%	2.3%	2.5%
A	Y	W1	Y	4.19	4.51	42.7%	46.9%	2.6%	4.9%
A	Y	W2	Y	4.13	4.50	42.4%	46.8%	2.5%	2.6%
A	Y	T	Y	7.37	4.85	73.5%	50.5%	2.2%	2.4%
B	N	W1	N		**0.85**		**2.7%**		3.6%
B	N	W2	N		**0.81**		**2.4%**		0.8%
B	N	T	N	4.63	4.24	48.6%	44.9%	2.4%	2.7%
B	N	W1	Y	2.88	**1.99**	28.1%	**15.9%**	3.0%	3.4%
B	N	W2	Y	2.85	**1.95**	27.5%	**15.1%**	3.0%	2.2%
B	N	T	Y	4.79	4.11	49.6%	42.7%	2.5%	2.6%
B	Y	W1	N		1.64		11.7%		3.8%
B	Y	W2	N		1.65		11.7%		3.6%
B	Y	T	N	4.64	4.24	48.6%	44.9%	2.4%	2.7%
B	Y	W1	Y	2.89	2.42	28.1%	22.9%	3.0%	3.2%
B	Y	W2	Y	2.85	2.40	27.5%	22.4%	3.0%	3.3%
B	Y	T	Y	4.80	4.11	49.7%	42.7%	2.5%	2.6%

Table 3.9: Simulation study results: selected samples under the alternate.

For unselected samples, the three treatments of allele frequency (P, population, E, estimated and S, sample) give equivalent results. Also, for selection scheme A (which is symmetrical and therefore does not bias allele frequency under an additive model) all three methods are more or less identical. Only selection scheme B, in both singletons and pairs, is asymmetrical and so distorts the sample allele frequency. In singletons without parents, this effect was slight (0.53 versus 0.50) whereas in pairs without parents this effect was larger (0.57 versus 0.50). For the samples with parents, the average discrepancy in allele frequency in selected samples was, by chance, larger for singletons (0.62 versus 0.50) and pairs (0.69 versus 0.50). The pattern of results

			LRT		p		VC	
Test	Freq	Par	$s=1$	$s=2$	$s=1$	$s=2$	$s=1$	$s=2$
W1	P	N		3.90	0.50	0.50		0.12
W1	E	N		3.90	0.51	0.51		0.15
W1	S	N		3.85	0.53	0.57		0.03
W2	P	N		3.89	0.50	0.50		0.08
W2	E	N		3.85	0.53	0.57		0.08
W2	S	N		3.85	0.53	0.57		0.08
T	P	N	19.61	17.35	0.50	0.50	0.11	0.11
T	E	N	12.53	4.69	0.51	0.51	0.11	0.13
T	S	N	8.22	1.07	0.53	0.57	0.05	0.01
W1	P	Y	11.59	7.92	0.50	0.50	0.12	0.11
W1	E	Y	11.60	7.78	0.50	0.55	0.12	0.15
W1	S	Y	11.68	7.77	0.62	0.69	0.07	0.06
W2	P	Y	10.88	7.60	0.50	0.50	0.12	0.10
W2	E	Y	11.02	7.87	0.56	0.63	0.12	0.10
W2	S	Y	11.50	8.28	0.62	0.69	0.12	0.10
T	P	Y	20.37	15.92	0.50	0.50	0.11	0.10
T	E	Y	17.11	8.19	0.50	0.50	0.11	0.11
T	S	Y	8.16	0.87	0.62	0.69	0.05	0.01

Table 3.10: Simulation study results: treatment of allele frequency.

is similar for samples with and without parental genotypes in any case. Table 3.10 shows the results for the different treatments, P, E and S under different analytic models for B selected samples. In all cases, the analysis is conditional on trait values (i.e. allele frequency cannot be estimated otherwise). The results show the LRT, the fixed or estimated value of allele frequency p and the total QTL variance component (VC) (i.e. which should be 0.1).

The power of the within tests W1 and W2 appears to be independent of allele frequency treatment. In contrast, the total association test T shows the expected pattern of results $P > E > S$: power is optimal when using the population allele frequency and worse when using the sample allele frequency. Estimating allele frequency provides intermediate power. The estimated allele frequencies under the W2 test are closer to the sample frequencies, whereas for both the W1 and T tests the estimated

values are closer to the population values than the sample values.

In general, it appears good advice to estimate the allele frequency when the true value is not known with any certainty (rather than use the sample allele frequency), when testing between-sibship effects in selected samples.

3.6 Other selection issues in association studies

This final section briefly explores selection issues in the context of three different association study designs. First, an approximation for the NCP of a standard quantitative trait association test in selected samples is demonstrated. Second, the properties of a two-stage association model are investigated by simulation. Third, in the context of DNA pooling, determination of pool threshold is discussed; in particular, a method for determining the optimal thresholds (for a given genetic model) of multiple (> 2) pools is presented.

3.6.1 Approximation based on change in variance

A sample selected to increase power will typically have a greater trait variance compared to the unselected sample it originated from. That is, for example, selecting only the 5% high and low scoring individuals will increase the average deviation from the mean score. Conversely, a poor selection strategy, say only selecting individuals scoring within 0.2 standard deviations from the mean, would lead a selected sample with a smaller variance than the unselected sample.

The change in variance from the unselected to selected sample should partially predict the efficiency of the design, therefore. If the unselected sample NCP, variance and sample size are λ, V and N respectively, then the expected NCP per sibship of the selected sample is

$$\frac{\lambda_S}{N_S} \approx \frac{V_S}{V} \cdot \frac{\lambda}{N}$$

if V_S and N_S are the variance and sample size of the selected sample. The accuracy of this approximation is considered firstly for samples of unrelated individuals, secondly for sibling pairs.

Unrelated individuals

Fifteen selected sampling schemes are applied to samples of 1200 unselected individuals, in which a 2% or a 10% QTL is present. Ten symmetric schemes extracted between 5% and 50% of the unselected sample, in 5% intervals, evenly from high and low ends of the distribution. Five asymmetrical schemes sampled between 5% to 50% of the unselected sample in 10% intervals, selecting from the top and bottom of the trait distribution in a 5:1 ratio (i.e. oversampling the high end).

Figure 3.7 shows the absolute NCP obtained by simulation: for each QTL size, two slopes are visible, corresponding to the symmetric versus the asymmetric schemes, i.e. the steeper, "more efficient" slope represents the symmetric schemes. The variance of the symmetric selected samples is also greater than variance of the asymmetric selected samples.

Figure 3.8 plots the increase in trait variance (proportional to the unselected sample variance) against the corresponding increase in NCP per individual, for all 15 selection schemes. All the points fall very close to the $y = x$ line at 45°, indicating that the approximation is a good one in this scenario.

Sibling pairs

For sibling pairs, it is necessary to consider separately the variance of the trait mean and absolute trait difference, which will correspond to the between and within sibship components of association respectively. Again, 15 selection schemes were applied: ten symmetric schemes selecting the 5% to 50% of sibling pairs maximally discordant for the trait; five schemes selecting the 10% to 50% of sibling pairs with the greatest sum

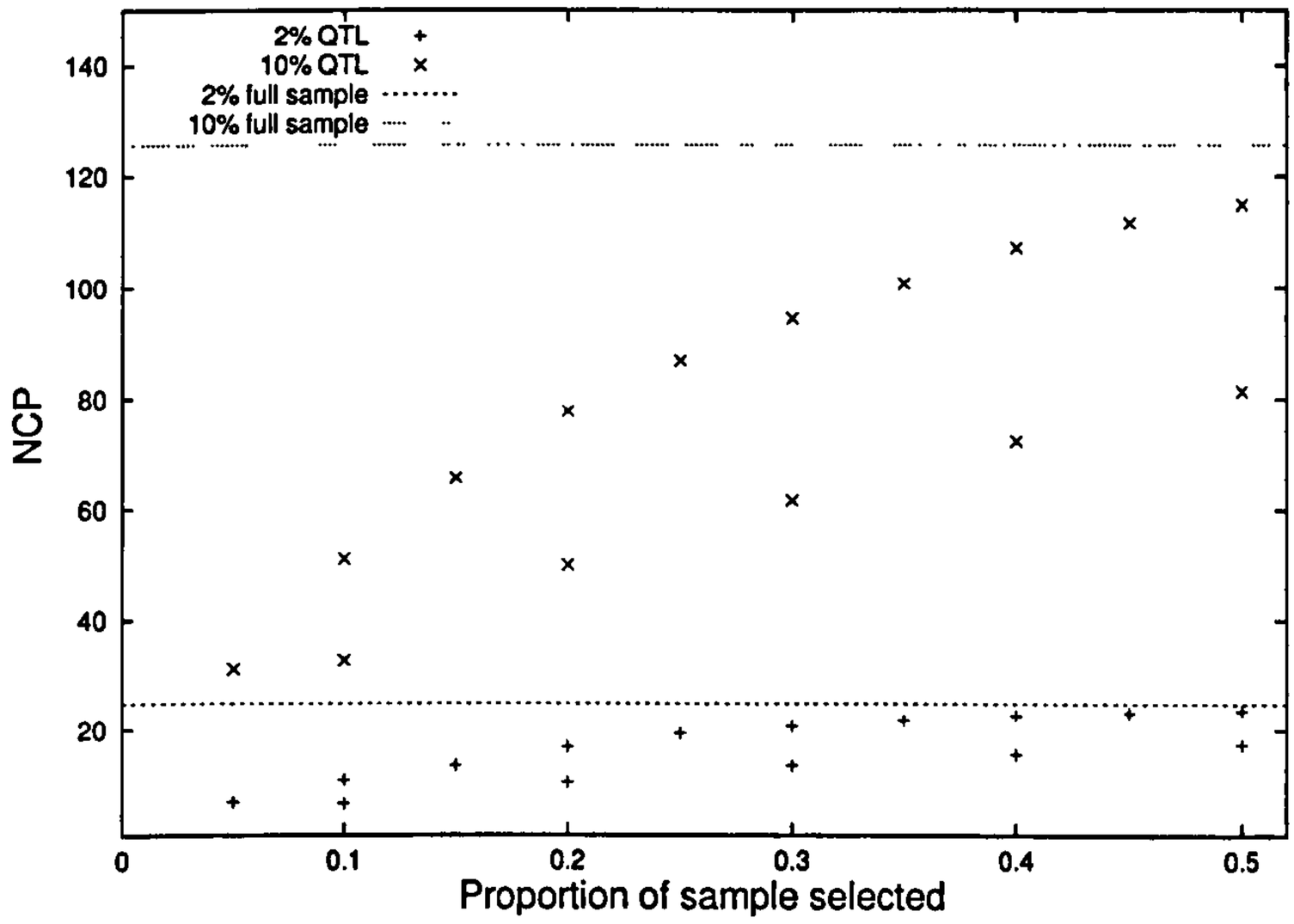


Figure 3.7: Average NCP in selected samples for additive QTL explaining 2 and 10% of the trait variance.

(i.e. concordant high pairs). The change of variance of the sibship trait difference should predict the change in NCP per sibship of the within test of association; likewise, the change in variance of the sum should predict the change in NCP per sibship of the between test. Both these predictions were found to hold: Figure 3.9 illustrates the relationship for the within test. The relationship is not quite as strong as for singletons (the 2% QTL case deviates from the expected slope, although the 10% QTL case does not).

Initial results suggest that the difference in variance between the unselected and selected samples can provide a reasonable guide to the efficiency of a given selected sampling scheme for singletons or pairs.

3.6.2 Two-stage association designs

This section considers a restricted problem in a different scenario, for a two-stage case-control association study (i.e. a binary disease trait). The basic parameters are the number of cases N_A and controls N_B and the allele frequency in cases p_A and

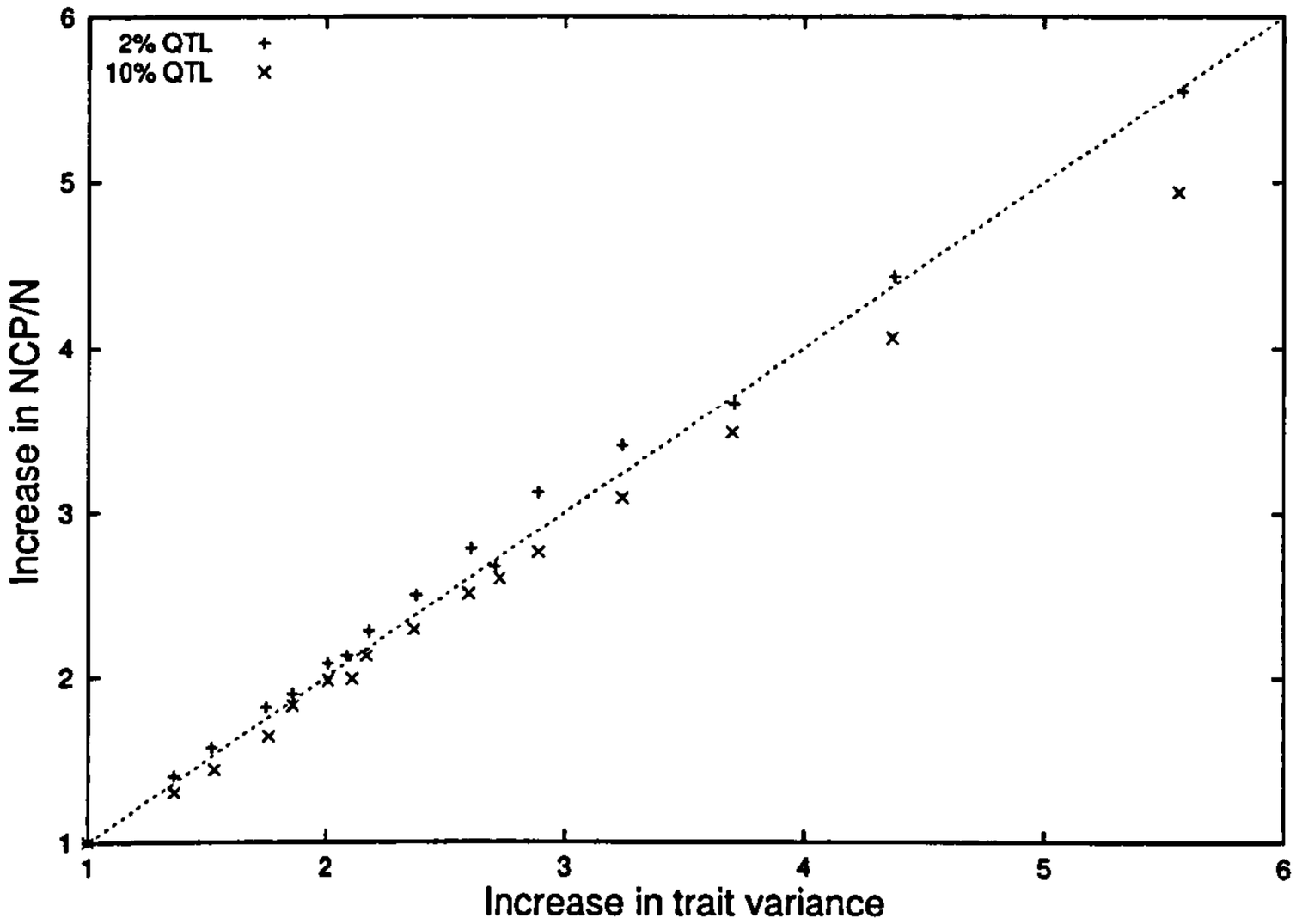


Figure 3.8: Approximation for NCP per individual in selected samples.

controls p_B. The two-stage selection procedure is defined by three further measures: the proportion of the sample to be genotyped in the first stage, r and the critical p-values for statistical significance in the first and second stages, α_1 and α_2.

The design is to genotype rN_A cases and rN_B controls in stage 1 at each marker. If that marker shows a significant result for a χ^2 test of gene–disease association, i.e. with a p-value less than α_1, then the remaining $(1-r)N_A$ cases and $(1-r)N_B$ controls are genotyped and a second χ^2 test statistic is calculated for the entire sample of cases and controls from both stages. This sequential testing procedure makes exact analytic solutions difficult, as stage 2 only occurs if stage 1 shows a significant result. Otherwise, the power would simply be the product of powers for the two stages.

In each replicate, the allele frequency in cases is calculated using the random cumulative binomial distribution function. Then a simple χ^2 test of independence is conducted on stage 1 data. If there is a significant result at stage 1, this process is repeated for stage 2, the data pooled and a further χ^2 test conducted.

Under the null (i.e. allele frequencies specified to be equal between cases and controls), the proportion of significant replicates estimates the Type I error rate; under

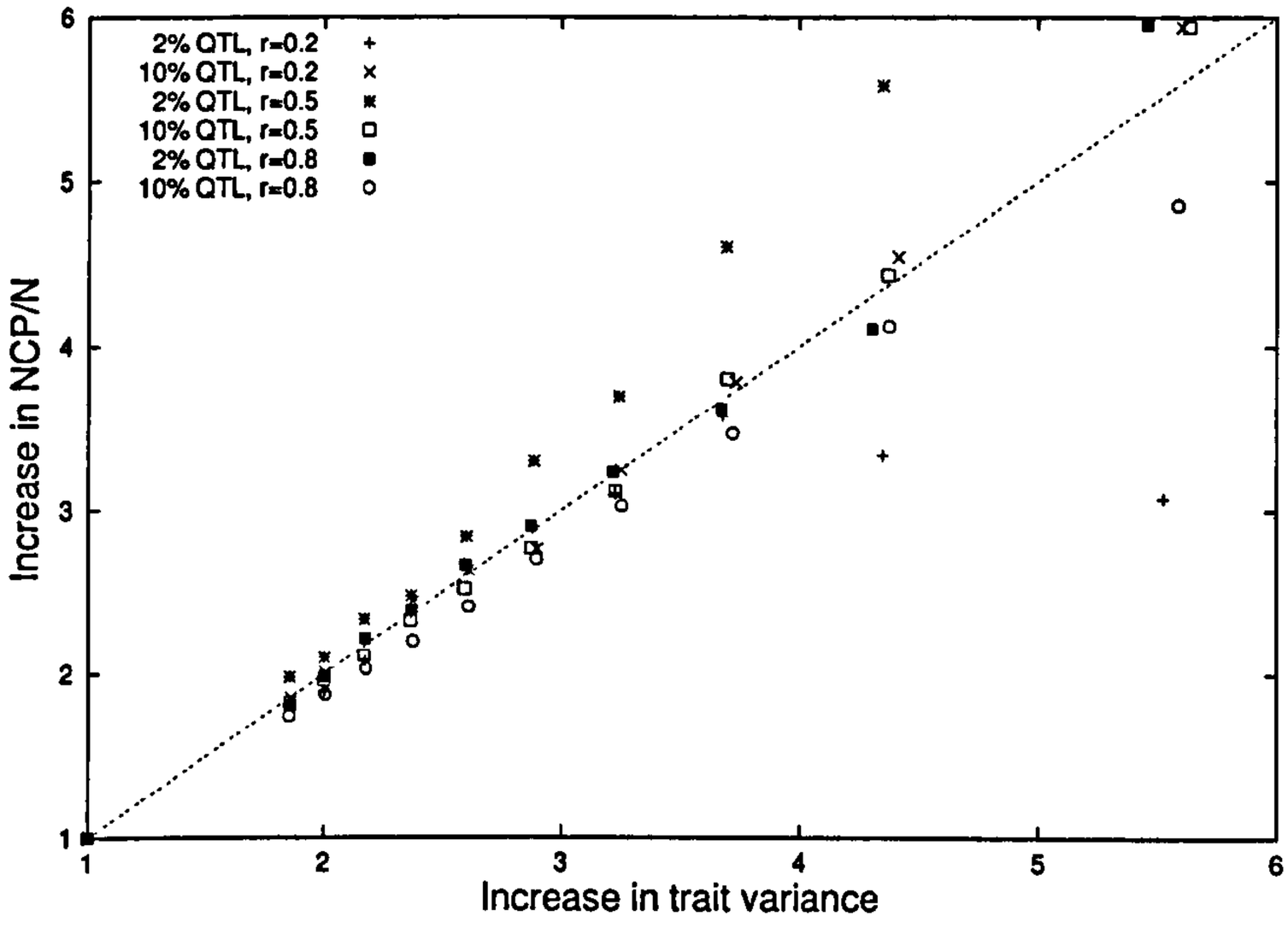


Figure 3.9: Approximation for NCP per sibling pair in selected samples.

the alternative hypothesis (i.e. allele frequencies specified to be different between cases and controls) the proportion of significant replicates gives the power. This procedure is implemented in the computer program `twostage`.

Results

To illustrate this procedure, consider the following example. For a diallelic QTL, the allele frequency was set at $p = 0.1$ in the population. Under the null, $p_A = p_B = 0.1$. Under the alternative, $p_A = 0.1429$ and $p_B = 0.09524$ where chosen such that with $N_A = N_B = 790$ one would have 80% power to detect an association.

The proportion of individuals to be genotyped in the first stage, r, was calculated to give a constant expected number of genotypes, no matter which Type I error rate was used in stage 1. That is, under the null, the expected proportion of genotypes g is $r + \alpha_1(1 - r)$. Therefore,

$$r = \frac{g(N_A + N_B) - \alpha_1(N_A + N_B)}{(N_A + N_B) - \alpha_1(N_A + N_B)}$$

For example, if one is prepared to genotype, on average, 40% of the sample, with $\alpha_1 = 0.05$, then $r = 0.368$. If, however, one wishes to use $\alpha_1 = 0.25$ then $r = 0.2$. Utilising the appropriate values of r, Table 3.11 represents the results for the above example. Roughly speaking, to obtain 80% power overall, one must genotype 60% of the sample on average. There is a complex, nonlinear relationship between α_1 and g in determining type I error rate and power. Further simulations are necessary to more properly characterise this situation.

		Overall Type I			Overall power		
g	$\alpha_1 =$	0.05	0.1	0.25	0.05	0.1	0.25
10		0.0006			0.0119		
20		0.0006	0.0005		0.1113	0.0568	
30		0.0005	0.0006	0.0005	0.2997	0.2423	0.0198
40		0.0007	0.0004	0.0005	0.5220	0.4827	0.2060
50		0.0007	0.0006	0.0009	0.7018	0.6942	0.4983
60		0.0009	0.0010	0.0011	0.8277	0.7267	0.9379
70		0.0009	0.0010	0.0010	0.8072	0.9294	0.8974

Table 3.11: Power calculation for two-stage association designs.

3.6.3 Sample selection in DNA pooling

DNA pooling is an efficient strategy for association analysis (Sham et al., 2002a). Typically, the allele frequency estimated for a pool of "cases" will be compared with the allele frequency estimated for a pool of "controls". In the case of a continuous trait, it is logical to compare the allele frequencies between a pool of "high scorers" and a pool of "low scorers". This necessarily involves choosing the trait thresholds that determine which individuals will go into the two pools. Forming one pool from extreme high scorers and another from extreme low scorers might appear to be desirable, in order to maximise the difference in pool allele frequencies. However, more extreme pools will consist of fewer individuals and so decreasing pool sample size will decrease power. An optimal strategy for DNA pooling will therefore involve finding the combination of effect size (allele frequency difference between pools) and sample

size which results in the greatest power to detect an association.

Power calculation for DNA pooling with two pools

Initially, consider the situation where two pools are formed from an unselected sample of unrelated individuals, on the basis of two thresholds, t_1 and t_2. That is, if $t_1 < t_2$, the first pool contains all individuals scoring $< t_1$; the second pool contains all individuals scoring $> t_2$. Fixing the total trait variance to unity and assuming a diallelic QTL, the total proportion of variance the QTL accounts for (σ^2_Q) is specified, the ratio of dominance to additive effects (z) and the allele frequency (p). The additive genetic value is

$$a = \sqrt{\frac{\sigma^2_Q}{2pq(1+z(q-p))^2 + (2pqz)^2}}$$

and $d = za$ so $\sigma^2_A = 2pq(a + d(q-p))^2$ and $\sigma^2_D = (2pqd)^2$. Residual variance $\sigma^2_R = 1 - (\sigma^2_A + \sigma^2_D)$. For QTL genotypes 11, 12 and 22, the genotypic values μ_{11}, μ_{12} and μ_{22} are specified a, d and $-a$ and then mean-centred by the mean, $a(p-q) + (2pqd)$. Genotype frequencies are p^2, $2pq$, q^2.

The calculation can be specified not for the QTL but instead for a marker in linkage disequilibrium with the QTL. This involves specifying a measure of linkage disequilibrium D' (D-prime). If marker allele frequencies are specified also, the haplotype frequencies can be calculated and used in subsequent calculations. Only the case where the marker is the QTL will be discussed in this Chapter however.

For each pool, the calculation of allele frequency proceeds in two stages: first the area under the three genotype distributions is calculated, as shown in Figure 3.10. For pool i, defined by lower and upper thresholds t_{il} and t_{iu}, the total area under the curve up to t_{iu} is

$$F_{iu} = p^2\Phi\left(\frac{t_{iu}-\mu_{11}}{\sigma_R}\right) + 2pq\Phi\left(\frac{t_{iu}-\mu_{12}}{\sigma_R}\right) + q^2\Phi\left(\frac{t_{iu}-\mu_{22}}{\sigma_R}\right)$$

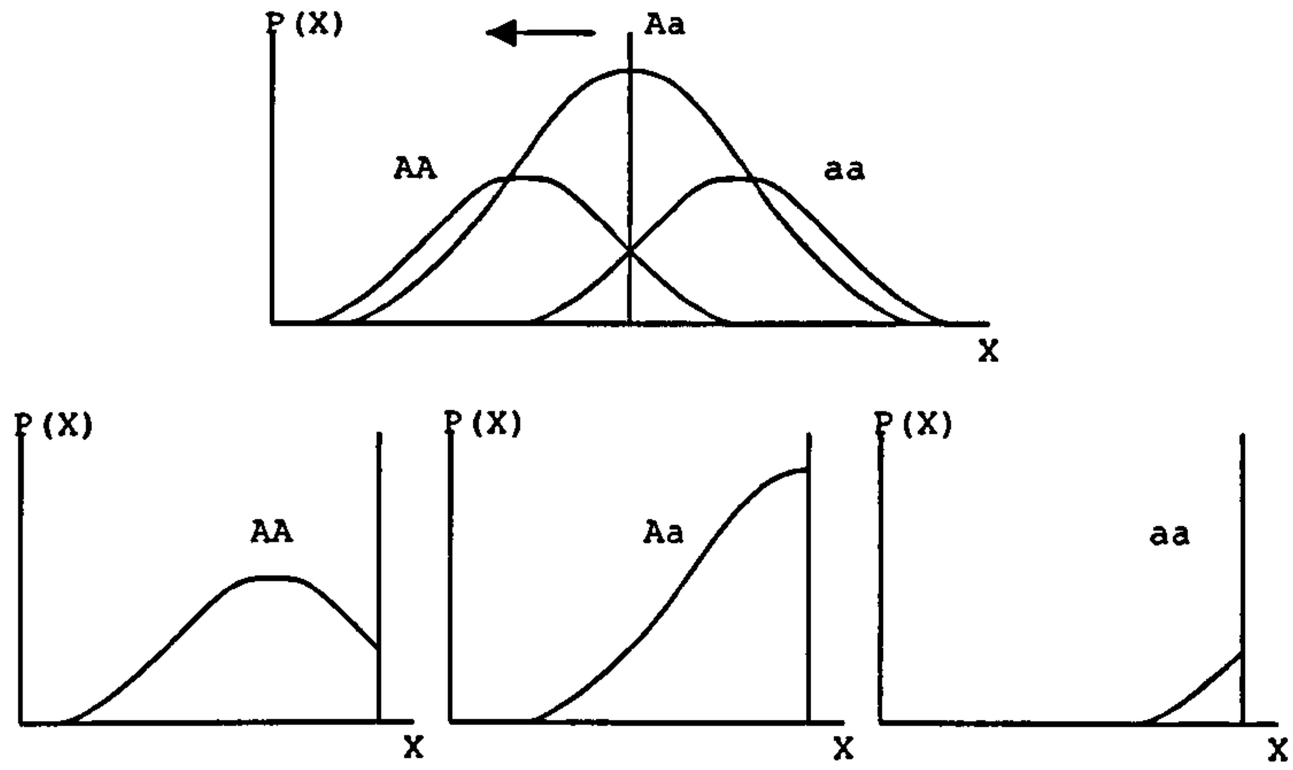


Figure 3.10: Power calculation for DNA pooling.

using the cumulative normal distribution function Φ. A similar calculation gives the area under the curve up to the lower threshold, F_{il}. To calculate the frequency of allele '1',

$$Q_{iu} = p^2\Phi\left(\frac{t_{iu} - \mu_{11}}{\sigma_R}\right) + pq\Phi\left(\frac{t_{iu} - \mu_{12}}{\sigma_R}\right)$$

and similarly for Q_{il}; the pool allele frequency is then $P_i = \frac{Q_{iu} - Q_{il}}{F_{iu} - F_{il}}$.

Assuming a single sample size for the unselected population, pool size will depend on the thresholds. That is, assuming the trait is normally distributed, the cumulative normal distribution function gives the proportion of the sample in the pool. The pool sample sizes can then be used to construct a 2×2 table (allele by pool), which gives a χ^2 test of pool-marker association.

For example, consider an additive diallelic QTL that explains 5% of the trait variance and has an allele frequency of 0.5. If a low and a high pool are selected using thresholds of -2 and 2 standard deviations above the mean, then assuming an unselected sample size of 15,000 would give approximately 345 individuals in each pool. Specifying the lower threshold for the lower pool as -8 standard deviations, and the upper threshold for the higher pool as 8 standard deviation is effectively equivalent to specifying simply '< 2' and '> 2' as the thresholds. This results in

expected pool allele frequencies of 0.32 and 0.68 for low and high pools respectively. The expected χ^2 statistic is 184.94 – highly significant for a single degree of freedom. This calculation is implemented as part of GPC.

Determining optimal threshold placement

For a given QTL model, the threshold values can be treated as free parameters and the expected χ^2 statistic maximised using numerical optimisation techniques (Nelder and Mead, 1965). In this way, the optimal placement of thresholds can be determined, conditional on the QTL model.

Table 3.12 gives the optimal thresholds assuming a symmetric design, i.e. $t_1 = -t_2$. For an additive QTL with equal allele frequencies of three different effect sizes (1%, 5% and 10%) the optimal threshold value was near 0.58 standard deviations from the mean. This corresponds to selecting approximately the top and bottom 27 – 28% of the sample to form high and low pools. This result is relatively invariant to the precise QTL model (i.e. rare allele, dominance).

	QTL variance		
	1%	5%	10%
t_1	0.583	0.583	0.561
Pool N	419	420	431
High pool p	0.54	0.60	0.63
Low pool p	0.46	0.40	0.37
χ^2	12.2	61.2	123.8

Table 3.12: Optimal threshold values for a two pool design: additive QTL with equal allele frequencies.

Considering a number of models, two main conclusions emerge: first, the average test statistic obtained from DNA pooling compared to standard individual genotyping association analysis of the entire unselected sample was typically around 60-70%. Second, the optimal pooling fraction under most models was approximately 27%. That is, selecting just over the highest quarter and lowest quarter of a sample to form two pools affords the most efficient pooling design.

In fact, these results for two pools have previously been demonstrated both analytically (Bader et al., 2002) and by numerical methods (Jawaid et al., 2002), and extended to include other factors such as error in pool allele frequency estimation and the impact of pool size on this error. The rest of this section considers a different extension, to designs with more than two pools as well as a corresponding method of analysis.

Multiple (> 2) pools

If pools are being constructed by grouping individuals scoring between certain thresholds, there is the possibility of incorporating more than two pools into a single analysis. For a continuous trait, it is possible to re-frame DNA pooling analysis within the maximum-likelihood variance-components framework. By anchoring the observed pool allele frequencies onto the trait distribution, with three or more pools additive and dominance components of variance at the test locus can be estimated. The method gives equivalent results to the standard approach with only two pools. Additional pools will increase the power of the association test, although the extent of the increase depends on the QTL model. It is also possible to estimate optimal thresholds given a set QTL model, adopting a similar logic to the 2 pool case, although exploring these issues is beyond the scope of the present section.

A maximum-likelihood method is used to estimate the parameters of the QTL model given the observed pooling data. For a standardised trait, the input data are, for each pool, the upper and lower pool thresholds, the number of individuals and the observed allele frequency. The QTL model parameters are additive genetic value a, dominance deviation d and population allele frequency p. Typically, all three parameters are estimated under the alternate hypothesis, whereas only p is estimated (with $a = d = 0$) under the null. Twice the difference in the log-likelihood under null and alternate models gives the likelihood ratio test for a QTL effect, which will have

either 1 or 2 degrees of freedom (depending if d was free in the alternate).

The data are the proportions of '1' alleles in pool i, which consists of N_i individuals. Assuming random mating, the probability of observing G_i '1' alleles of the $2N_i$ total alleles is given by the binomial distribution. Therefore the likelihood of the observed pool is

$$L(G_i|f_i) = \frac{G_i}{G_i!(2N_i - G_i)!} f_i^{G_i}(1 - f_i)^{2N_i - G_i}$$

where f_i is the pool i allele frequency. As shown in the previous section, this parameter is a function of the thresholds and the QTL model (a, d and p). That is $f_i = \frac{Q_{iu} - Q_{il}}{F_{iu} - F_{il}}$ where Q and F are calculated as before. For n pools, the sample log-likelihood is then $\sum_{i=1}^{n} \ln L(G_i|f_i)$. This likelihood can be maximised in order to obtain MLEs of the QTL model parameters.

The advantage of this method is that it can handle any number of pools (that may be partially or completely overlapping) and it directly estimates the genetic parameters of interest at the test locus (including the ability to estimate dominance effects if more than two pools are analysed). When the unselected sample size is very large (e.g. 15,000) then the utility of using multiple pools is apparent, as absolute pool size is realistically restricted to less than 1,000 individuals for molecular reasons (Sham et al., 2002a). This method is implemented in the computer program `mpool`.

Results

Firstly, consider a simple two pool case, reported in Table 3.13, with an additive diallelic QTL with equal allele frequencies explaining 1% of the trait variance.

In all cases, the design is symmetric (i.e. $t_1 = -t_2$). The $E(\chi^2)$ column gives the expected test statistics based on a conventional contingency table test. The next two columns give the ML estimates of QTL variance and allele frequency under the alternate – they appear to be unbiased. The next column gives the ML estimate of p with $a = d = 0$ fixed. The final column gives the likelihood ratio test statistics,

		H_A		H_0	
t_1	$E(\chi^2)$	σ^2_A	p	p	$-2(\ln L_A - \ln L_0)$
0	95.8	0.011	0.500	0.500	108.0
0.5	121.4	0.010	0.500	0.500	119.2
0.75	122.0	0.011	0.500	0.500	138.2
1	110.3	0.009	0.500	0.501	94.9
1.5	75.2	0.011	0.500	0.500	79.0
1.75	55.3	0.011	0.500	0.500	61.7
2	38.5	0.009	0.500	0.500	35.5
2.25	23.9	0.010	0.500	0.500	23.5
2.5	14.1	0.010	0.500	0.500	14.4
3	3.14	0.009	0.500	0.500	2.9

Table 3.13: Analysis results for 2 pools: 1% additive effects QTL, equal allele frequencies, 2 pools.

which are comparable to their expected values (note: only a single simulation was conducted here, for purely illustrative purposes).

Extending the illustration to include dominance effects and multiple pools, consider a QTL explaining 5% of the trait variance, with $p = 0.9$ and a dominance-to-additive effects ratio (z) of 1. The expected variance components are 0.0091 and 0.0409 for σ^2_A and σ^2_D respectively. The unselected population consists of 15,000 individuals, and three scenarios will be considered: 2 pools of 500 individuals, 3 pools of 333 individuals and 5 pools of 200 individuals. All thresholds are selected to be symmetrical about the mean (i.e. the 'odd' pool in the 3 and 5 pools cases includes the mean). The expected pool-specific allele frequencies are calculated (given the QTL model and the pool thresholds). These values are then entered as the 'observed' pool allele frequencies in `mpool`, which attempts to recover the original variance components. The thresholds and expected allele frequencies are given in Table 3.14.

Table 3.15 gives the ML parameter estimates and likelihood ratio tests for the different pooling scenarios (including and excluding dominance for the > 2 pool scenarios). In the two pool case, the test value of 105.151 is the χ^2 obtained from conventional pooling analysis. Note that the QTL variance and allele frequency have been under-estimated however.

	N	t_l	t_u	G_i
2 pools				
Pool 1	500	1.834	7.000	0.909
Pool 2	500	-7.000	-1.834	0.737
3 pools				
Pool 1	333	2.01	7.000	0.909
Pool 2	333	-0.028	0.028	0.908
Pool 3	333	-7.000	-2.010	0.689
5 pools				
Pool 1	200	2.216	7.000	0.909
Pool 2	200	0.643	0.685	0.909
Pool 3	200	-0.017	0.017	0.908
Pool 4	200	-0.685	-0.643	0.906
Pool 5	200	-7.000	-2.216	0.618

Table 3.14: Thresholds for multiple pools.

In the 3 pool scenarios, without modelling dominance the χ^2 value is greater than for two pools, but the parameter estimates are distorted (as the model is misspecified when dominance is not included). When a dominance effect is allowed for, then the three parameters are recovered almost exactly, resulting in an even higher test statistic (147.566). Although this now has 2 degrees of freedom, the specific test for dominance is significant also (around 12.363 with 1 degree of freedom). The results for the 5 pool scenario show a similar trend, with an even higher χ^2 value. In other words, by using more pools based on the same number of individuals the ability power of the QTL association test has been increased.

Of course, using 5 instead of 2 pools increases the amount of genotyping necessary by 2.5 times – the increase-factor in the χ^2 is only around 1.66. For three pools a genotyping increase of 1.5 times corresponds to a χ^2 increase of approximately 1.4 times. These increases in the χ^2 statistic are specific to the QTL model, and are likely to be much less under more straightforward additive models.

Multiple pools might be desirable for other reasons, however. As mentioned, there are absolute limits on pool size, so if the unselected sample is very large, it may be

	H_A			H_0	
	σ^2_A	σ^2_D	p	p	$-2(\ln L_A - \ln L_0)$
2 pools	0.022	—	0.838	0.823	105.151
3 pools	0.033	—	0.852	0.835	122.842
3 pools	0.009	0.041	0.899	0.835	147.566
5 pools	0.045	—	0.864	0.849	130.077
5 pools	0.009	0.041	0.900	0.849	175.904

Table 3.15: Analysis results for multiple pools: 5% QTL including dominance variance.

necessary to make several smaller pools. Additionally, these illustrations do not take into account the potentially important effect of allele frequency estimation error – in the presence of relatively high error variance, using multiple pools will possibly show a greater advantage.

3.7 Summary

In summary, this Chapter has described a novel method of sample selection of sibships for the Fulker association test and a method of association analysis that is robust in selected samples. Additionally, several subsidiary issues were briefly discussed: an approximation for selected sample NCPs, calculating power for two-stage association designs and DNA pooling analysis with more than two pools.

All of the above work has assumed that the test locus is the QTL itself, rather than in linkage disequilibrium with it. Abecasis et al. (2001a) considered the impact of sample selection on linkage disequilibrium mapping using sibships, finding that different selection schemes can sometimes exacerbate the attenuation in signal due to incomplete linkage disequilibrium. Examining the properties of the current method in the context of linkage disequilibrium mapping is an obvious next step. It is worth exploring whether the treatment of allele frequency during trait-conditional analysis would impact on this problem in any way.

An alternative approach to the analysis of selected samples is to generate empirical significance values by use of a permutation test, (e.g. Abecasis et al., 2000, im-

plemented in the QTDT computer program). A comparison of these methods would be of interest – it is possible that the conditional approach might be more powerful if the unselected model is correctly specified (e.g. unselected trait mean, variance and sibling correlation, and possibly allele frequency).

A further area of interest is the ease with which the conditional model can be extended to include other effects, such as covariates, epistasis and gene–environment interaction. Chapter 7 investigates the inclusion of covariates and gene–environment interaction in a conditional model; Chapter 8 considers the case for epistasis.

Part II

Complex Effects

Chapter 4

Gene × environment interaction in twin analysis

Gene–environment interaction is likely to be a common and important source of variation for complex behavioural traits. Often conceptualised as the genetic control of sensitivity to the environment, it can be incorporated in variance components twin analyses by partitioning genetic effects into a mean part, which is independent of the environment, and a part that is a linear function of the environment. The model allows for one or more environmental moderator variables (that possibly interact with each other) that may i) *be continuous or binary* ii) *differ between twins within a pair* iii) *interact with residual environmental as well as genetic effects* iv) *have nonlinear moderating properties* v) *show scalar (different magnitudes) or qualitative (different genes) interactions* vi) *be correlated with genetic effects acting upon the trait, to allow for a test of gene–environment interaction in the presence of gene–environment correlation. Aspects and applications of a class of models are explored by simulation, in the context of both individual differences twin analysis and, in Chapter 7, sib-pair quantitative trait locus linkage analysis. As well as elucidating environmental pathways, consideration of gene–environment interaction in quantitative and molecular*

studies will potentially direct and enhance gene-mapping efforts.

4.1 Overview

4.1.1 Current aims

The initial framework for the analysis of $G \times E$ in the context of the twin study has existed for some time. For example, Martin et al (1987) describe a model to handle continuous moderator variables and interactions between both latent and measured genetic and environmental effects, as well as documenting the power of such tests. This Chapter aims to extend this seminal work and to provide more comprehensive simulation results. The Chapter is split into five main sections: (1) the basic $G \times E$ model involving a continuous moderator variable that can interact with latent genetic and environmental effects (2) nonlinear $G \times E$ using a quadratic approximation (3) $G \times E$ in the presence of r_{GE} (4) scalar (different magnitudes) and qualitative (different genes) interactions (5) the impact of distributional factors on $G \times E$ analysis.

Some notation is introduced in order to clarify different moderating effects. Standard $G \times E$ will be called $A \times M$: the G is replaced by A to refer specifically to additive genetic effects; E is replaced by M (moderator), to distinguish it from the latent nonshared twin environment. Other types of interaction are $C \times M$ and $E \times M$, where the latent shared and nonshared environments, respectively, interact with a measured moderator and, in the companion paper, $Q \times M$ interaction, where a specific QTL interacts with a moderator. The term $G \times E$ will still be used to refer to the whole class of these effects.

4.2 $G \times E$ with continuous moderator variables

A naive treatment of continuous moderation might proceed as follows: stratify the sample into a number of groups on the basis of the moderator, calculate heritability within each strata, equate parameters across strata or test for a linear trend in heritability across strata. There are, however, several problems with such an approach. Firstly, the stratification procedure will effectively reduce the sample size, especially if the moderator is not shared between twins. Secondly, the use of heritability essentially assumes equal variance across strata, whereas what is of interest is whether the absolute magnitude of genetic effects changes, not only the proportion. Thirdly, although it is logical to initially assume a linear $G \times E$ interaction, this linearity should be at the level of *effect* rather than the level of *variance component*, as variance is a second-order statistic.

Consider the basic biometrical model for a hypothetical additive diallelic trait locus, with additive genetic value a and increaser allele frequency p. The locus' contribution to the variance, $2p(1-p)a^2$, is a function of both the square of magnitude of effect and how common it is. A linear $A \times M$ interaction implies that the additive genetic value is a linear function of the moderator M, namely $a + \beta M$ where β is an unknown parameter to be estimated. If β is significantly non-zero, this is evidence of a $A \times M$ interaction. The contribution to the variance is $2p(1-p)(a+\beta M)^2$, indicating that variance is a quadratic function of the moderator under linear interaction. Figure 4.1 illustrates a linear interaction effect for a single hypothetical QTL.

This hypothetical QTL model directly translates into the twin model. Path coefficients represent the magnitude of effect and so we express the path coefficients as linear functions of a moderator. In other words, the additive genetic path coefficient is no longer a, it is now $a + \beta_X M$. Therefore, if β_X is significantly non-zero, this represents an $A \times M$ interaction. The moderator may be obligatorily shared or it can be specified separately for each twin (e.g. age and parental income are obligatorily

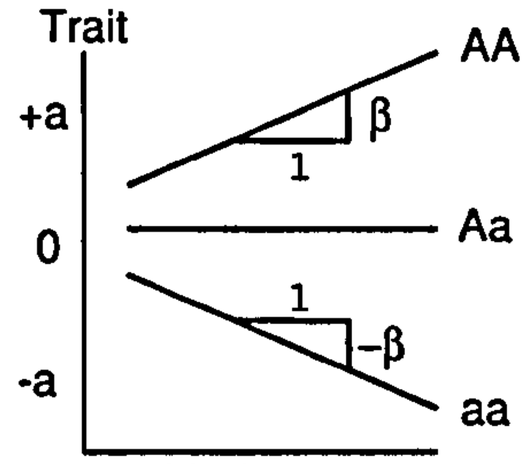


Figure 4.1: The biometrical model incorporating linear $A \times M$ interaction; the coefficient β assesses the extent of interaction.

shared; weight and exposure to violence are not). Binary moderators can be coded as '0/1', in which case the model reduces to the standard 'stratify by environment' approach.

Any variable which has a moderating, or interactive, effect on a trait may also have a mediating, or main, effect. Therefore, the moderator can also be entered in the means model, where the parameter β_M represents the standard phenotypic regression coefficient. Additionally, we allow for $C \times M$ and $E \times M$ interaction: that is, of the measured moderator with either the residual latent shared or nonshared environmental variables, assessed by β_Y and β_Z respectively.

For each twin pair conditional on the twins' moderator M, the expected trait mean for twin i is $\mu + \beta_M M_i$ and the expected trait variance is

$$Var(T_i) = (a + \beta_X M_i)^2 + (c + \beta_Y M_i)^2 + (e + \beta_Z M_i)^2$$

for $i = 1, 2$. The expected MZ covariance is

$$\mathrm{Cov}_{MZ}(T_1, T_2) = (a + \beta_X M_1)(a + \beta_X M_2) + (c + \beta_Y M_1)(c + \beta_Y M_2)$$

the expected DZ covariance is

$$\mathrm{Cov}_{DZ}(T_1, T_2) = 0.5(a + \beta_X M_1)(a + \beta_X M_2) + (c + \beta_Y M_1)(c + \beta_Y M_2).$$

SD from mean	$(a+\beta_X M)$	$(a+\beta_X M)^2$	$Var(T)$	h^2
-3	0.4	0.16	2.16	0.07
-2	0.6	0.36	2.36	0.15
-1	0.8	0.64	2.64	0.24
0	1.0	1.00	3.00	0.33
1	1.2	1.44	3.44	0.42
2	1.4	1.96	3.96	0.49
3	1.6	2.56	4.56	0.56

Table 4.1: An $A \times M$ interaction: additive genetic variance and heritability is tabulated against different values of the moderating variable. Parameter values are $a = c = e = 1$ and $\beta_X = 0.2$, $\beta_Y = \beta_Z = 0$.

This is equivalent to the model used by Martin et al (1987), in which, for example, variance due to additive genetic effects and $G \times E$ is $a^2 \cdot (1+\beta M)^2$ (i.e. in the current formulation, their interaction coefficient is β_X/a).

Seven parameters (unmoderated components a, c and e; moderated components β_X, β_Y and β_Z; main effect β_M) are now estimated under the full model, $ACE-XYZ-M$. Figure 4.2 shows a partial path diagram representing the $ACE-XYZ-M$ model. The best-fitting model can be obtained by successively dropping either moderating, main effect and/or unmoderated components. Assuming that at least one moderated parameter remains estimated in the model, the results must be considered in the context of a sensible range of moderator variable values. The expected variance components representing additive genetic, shared environmental and nonshared environmental effects can be plotted as a function of M. For example, the additive genetic component is $(a+\beta_X M)^2$ for a sensible range of M. Clearly, to extrapolate beyond the range of M observed in the data could be misleading: an approach to a clearer visualisation of moderated variance components is outlined further below. Table 4.1 presents a simple example of calculating the additive genetic variance and heritability given parameter values for the $ACE-XYZ-M$ model – this indicates what an interaction of $\beta_X = 0.2$ (a value subsequently used in many of the simulations) actually 'looks like'.

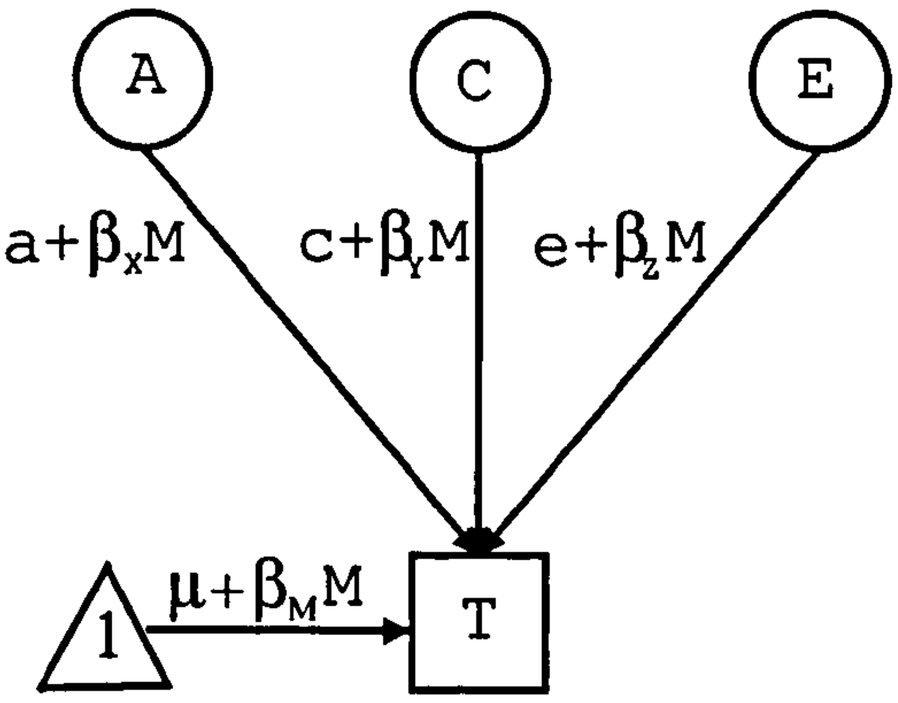


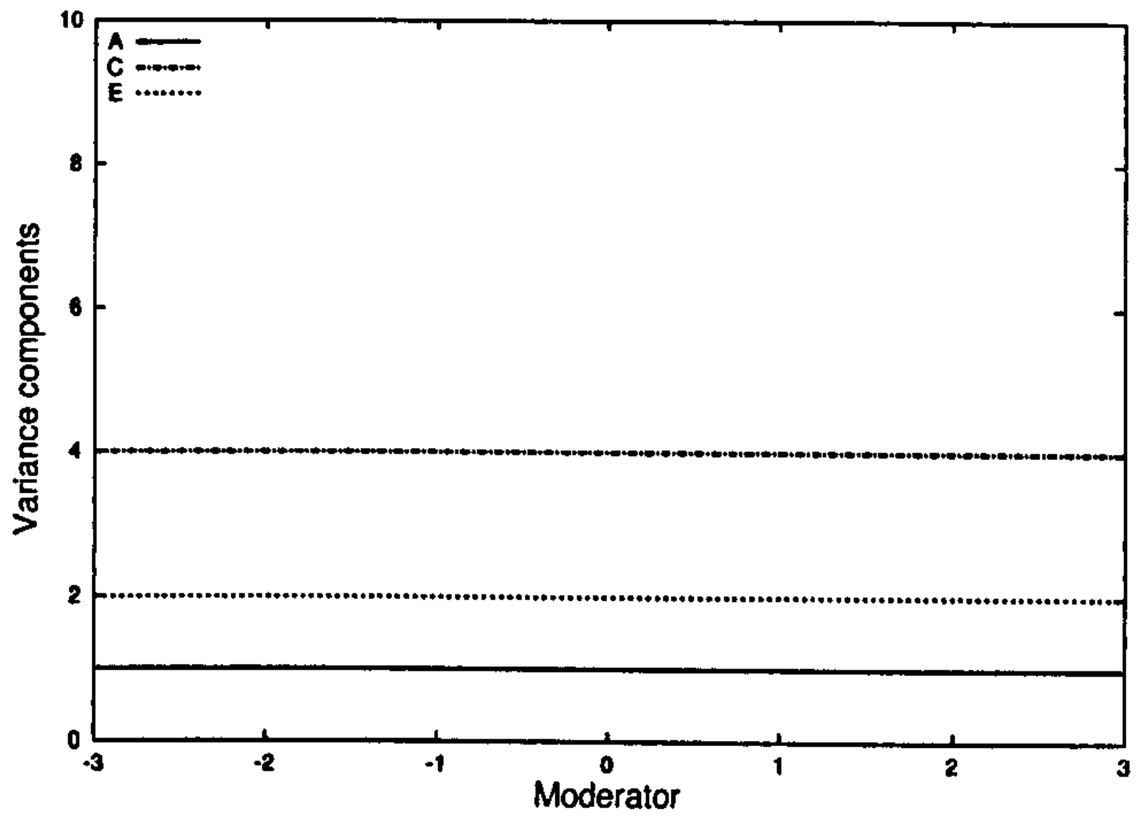
Figure 4.2: Partial path diagram for the $ACE - XYZ - M$ model, shown for one twin only. Latent variables have unit variance.

4.2.1 An example

A normally-distributed trait was simulated with A, C and E components representing 25%, 25% and 50% of the trait variance respectively. In addition, an obligatorily shared moderator variable was created with (1) a substantial main effect on the trait and (2) a marked interactive effect on the A component of the trait. The $A \times M$ was such that genetic effects were attenuated at intermediate values of the moderator but exaggerated at extreme high or extreme low values.

Fitting the various models starting with the full $ACE-XYZ-M$ model, the best-fitting model was the $ACE-X-M$ model, which correctly represents the simulation procedure described above. Figure 4.3 represents the variance components under the basic ACE model (i.e. equivalent to looking only at the trait and completely ignoring the moderator) and the best-fit model $ACE - X - M$. The signatures of mediation and moderation are clearly visible. Note that under the ACE model the C component is much greater than the A component, even though both residual components were simulated to account for 25% of the variance, because C includes the variance due to the main effect of the moderator (the moderator was obligatorily shared between twins). When the main effect is explicitly accounted for by the M component in the best-fitting $ACE - X - M$ model, the C component drops to the appropriate

a)



b)

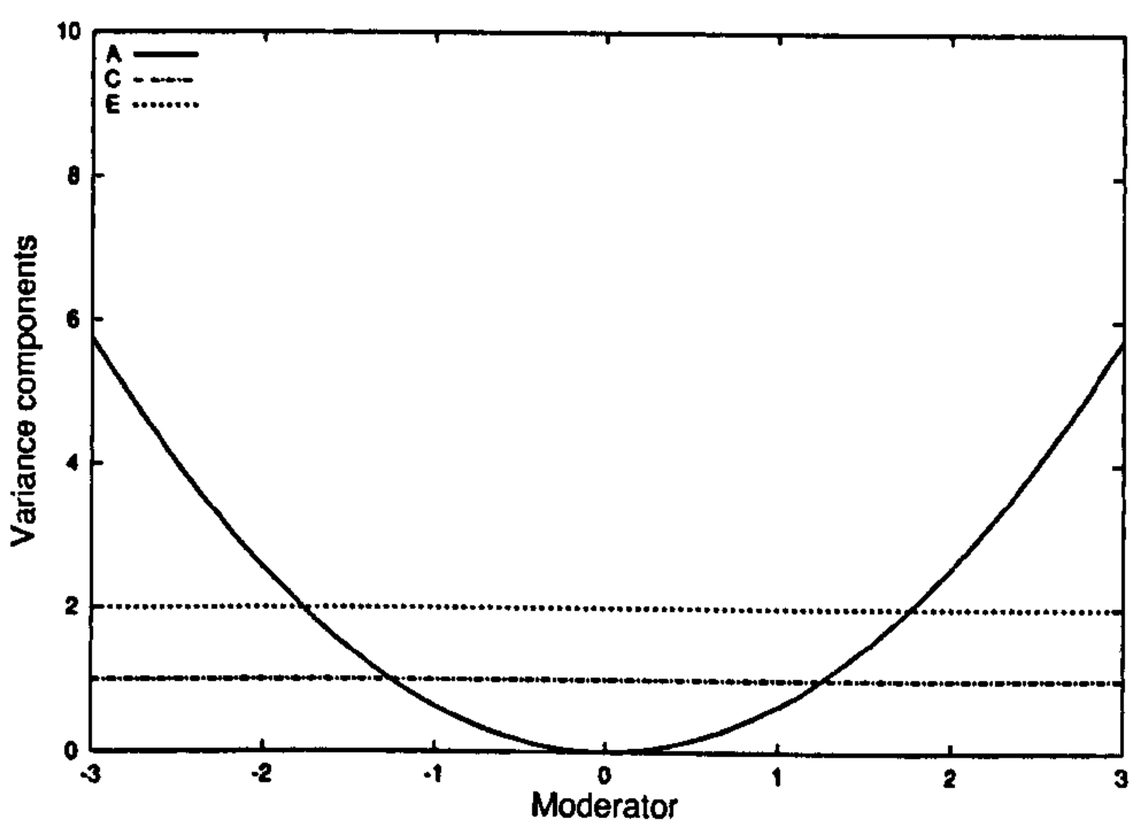


Figure 4.3: Modelling moderating and main effects: a) ACE model b) $ACE-X-M$ model.

level (i.e. half of E). As the β_X parameter is nonzero in the best-fitting model, we observe that the additive genetic variance varies as a function of the moderator, in a way which directly corresponds to the simulated properties (i.e. no genetic effects at intermediate levels of the moderator, exaggerated genetic effects at extreme values of the moderator).

If one were to standardise the variance components (for example, by plotting $(a+\beta_X M)^2 / ((a+\beta_X M)^2 + (c+\beta_Y M)^2 + (e+\beta_Z M)^2)$ the results will indicate proportions of variance. In the current example, genetic influences increase at extreme values of the moderator whilst environmental effects are constant. Proportionally, however, the environmental effects necessarily get smaller at extreme values, relative to genetic effects, as illustrated in Figure 4.4 which plots the expected variance and twin covariances as well as the standardised variance components for the current example. Arguably, this is somewhat misleading, and plotting only unstandardised results is encouraged.

4.2.2 Further simulations

A more comprehensive set of simulations was conducted in order to explore some of the properties of this model. A moderately large sample size of 500 MZ pairs and 500 DZ pairs was used under all models. Twin data were simulated for a continuous, normally-distributed trait and moderator variable. In all cases, the unmoderated parameter values were set at $a = c = e = 1$ (to give a variance of 3 excluding moderating and main effects). Table 4.2 gives the average parameter estimates and fit statistics for several conditions. Data were simulated under three true model conditions; each condition was replicated 50 times; each replicate was analysed under 8 nested models. The three true models were $ACE-X$, $ACE-Y$ and $ACE-Z$, representing $A \times M$, $C \times M$ and $E \times M$ interactions. The β interaction coefficients were either set at 0 (if not in the model) or 0.2. In all cases the moderator variable

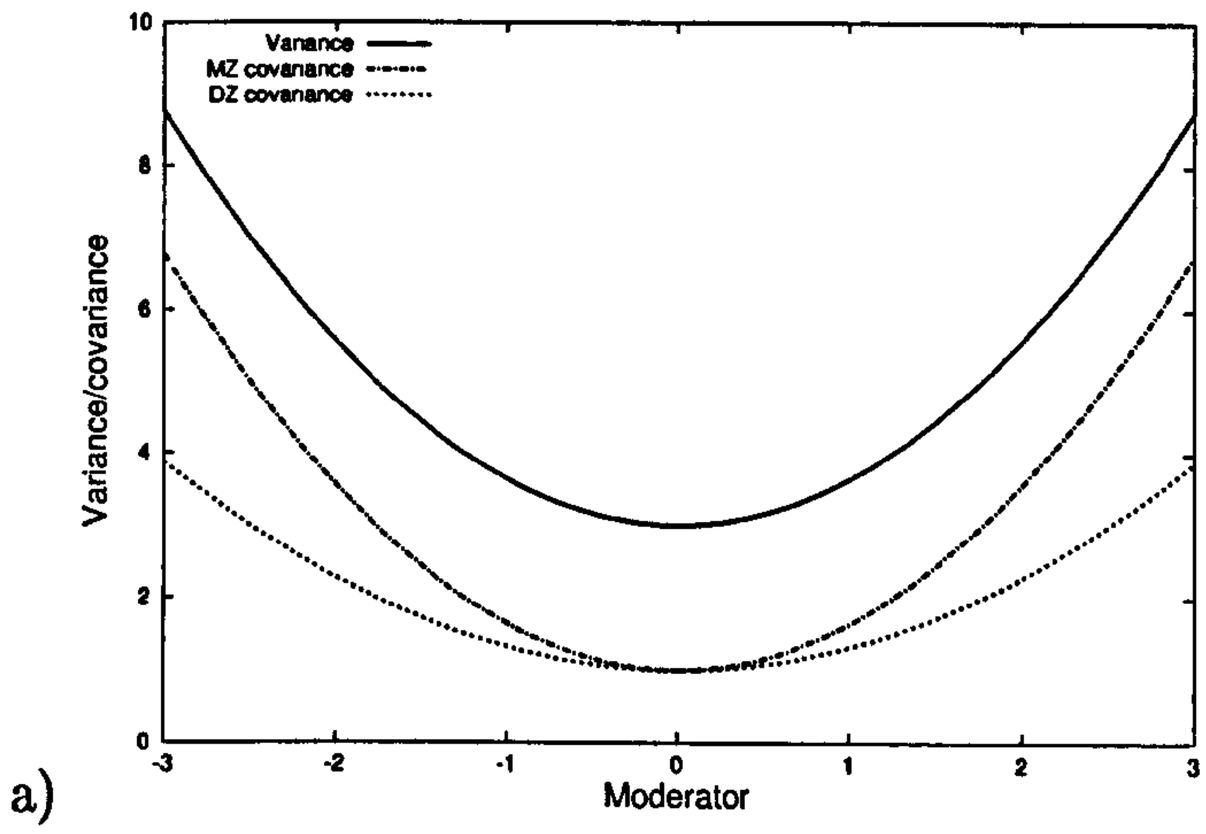


Figure 4.4: The impact of standardisation: a) the expected variance and twin covariances as a function of a moderator; in this example, the total trait variance differs across the range of the moderator; b) the standardised variance components.

True model	Analytic model	β_X	β_Y	β_Z	-2LL	AIC	% selected
$ACE-X$	$ACE-XYZ$	0.18	0.01	0.01	7433.95	-552.05	4
	$ACE-YZ$	.	0.16	0.04	7438.00	-550.00	6
	$ACE-XZ$	0.20	.	0.00	7434.88	-553.13	14
	$ACE-XY$	0.20	0.00	.	7435.08	-552.92	8
	$ACE-X$	0.20	.	.	7435.99	-554.01	62
	$ACE-Y$	.	0.18	.	7440.42	-549.58	6
	$ACE-Z$	.	.	0.07	7447.48	-542.52	0
	ACE	.	.		7455.44	-536.56	0
$ACE-Y$	$ACE-XYZ$	0.05	0.16	0.00	7436.76	-549.24	2
	$ACE-YZ$	.	0.21	0.00	7437.95	-550.05	10
	$ACE-XZ$	0.20	.	-0.02	7440.31	-547.69	6
	$ACE-XY$	0.04	0.16	.	7437.63	-550.37	6
	$ACE-X$	0.18	.	.	7441.45	-548.55	18
	$ACE-Y$	.	0.21	.	7438.85	-551.15	58
	$ACE-Z$	.	.	0.05	7453.59	-536.41	0
	ACE	.	.	.	7457.70	-534.30	0
$ACE-Z$	$ACE-XYZ$	-0.03	0.03	0.21	7415.41	-570.59	6
	$ACE-YZ$	.	0.00	0.21	7416.57	-571.43	12
	$ACE-XZ$	-0.01	.	0.21	7416.55	-571.45	4
	$ACE-XY$	0.26	-0.06	.	7452.27	-535.73	0
	$ACE-X$	0.21	.	.	7453.87	-536.13	0
	$ACE-Y$	.	0.15	.	7462.24	-527.76	0
	$ACE-Z$	.	.	0.21	7417.56	-572.44	78
	ACE	.	.	.	7472.70	-519.30	0

Table 4.2: Average parameter estimates and fit statistics for twin models of linear $A \times M$, $C \times M$ and $E \times M$ interaction.

was set to have a twin correlation (for both MZ and DZ twins) of 0.5. Results not shown here indicate a very similar pattern for other values, including 0 and 1 (also, the next set of simulations varies the moderator twin correlation). No main effects of the moderator are simulated or included in the model in this first set of simulations.

Table 4.2 shows the average best-fit parameter estimates as well as the averaged minus twice log-likelihood of the data and AIC index. The last column, "% selected" refers to the percentage of the 50 replicates that were selected from the 8 analytic models on the basis of AIC. The parameter estimates for the mean and unmoderated parameters are not shown: they were all very close to simulated values.

The full $ACE - XYZ$ model generally recovers the interaction parameters quite well. For example, for data simulated under the $ACE - X$ model, the average values of β_X, β_Y and β_Z are 0.18, 0.01 and 0.01 (i.e. true values 0.20, 0.00 and 0.00). In this case, the average $-2LL$ is 7433.95, whereas under the $ACE - X$ model it is 7435.99. The average difference, from dropping the Y and Z components is only 2.04, which is not significant for a χ^2_2 at $\alpha = 0.05$, suggesting that these terms can be dropped from the model.

There is clearly an issue of specificity here, however. For example, note that the β_Y coefficient of the $ACE-Y$ model is 0.18 even when the data were simulated under the $ACE-X$ model. That is, these interaction parameters are quite highly correlated (as are a and c), which can lead to some reduction in power to detect one in the presence of the other. However, on the basis of lowest AIC, the correct model was selected the majority of the time under all three conditions (62%, 58% and 78%). Typically, the second most selected model contained the correct interaction term, also. In none of the cases was the basic ACE model with no interaction terms supported.

Figure 4.5 illustrates the relationship between variance components and the expectations for twin variance and covariance as a function of a moderator variable. Quite different patterns of interaction can give rise to quite similar patterns of variance and covariance. Given that analysis moves from the observed variances and covariances to the inferred variance components, the relative indistinguishability of the models is unsurprising. Generally, however, the parameter estimates under the full model can serve as a guide to the true model. One strategy, therefore, is to plot the variance components using the parameter estimates of the full model – the general outline of this plot should not change greatly under nested submodels. This does highlight the danger of only testing for $A \times M$ interaction within this framework, however. For example, for data simulated under the $ACE - Y$ model, the average difference in -2LL between the $ACE - X$ and ACE models is 16.25, which is highly significant for

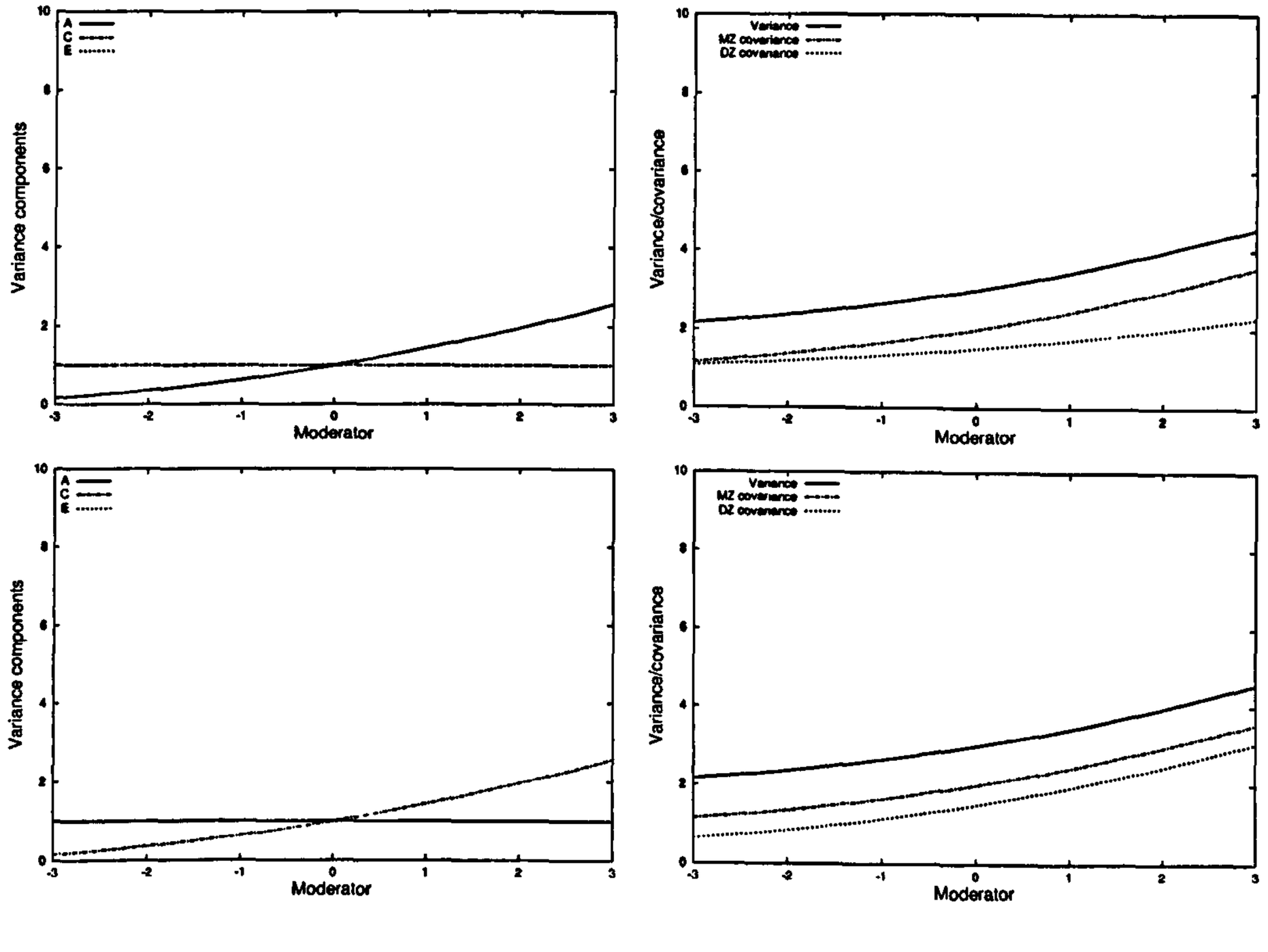


Figure 4.5: Relationship between variance components and expected variance, twin covariance. In both cases $a = c = e = 1$. In panel (a), $\beta_X = 0.2$ whilst $\beta_Y = \beta_Z = 0$. In panel (b), $\beta_Y = 0.2$ whilst $\beta_X = \beta_Z = 0$. Despite this marked difference in aetiology (lefthand graphs), the expected variances and covariances are remarkably similar (righthand graphs).

1 degree of freedom.

Several other properties are explored in the next set of simulations, the results of which are given in Table 4.3. Data were simulated under four models, and under different twin correlations for the moderator variable ($r = 0$, 0.5 and 1). The first model is simply the null model with no moderating or mediating effects. The second model represents combined interactive effects, with $\beta_X = \beta_Y = \beta_Z = 0.2$. The third model also includes a main effect, $\beta_M = 0.2$. The final model has two opposing interactive effects, $\beta_X = 0.2$ and $\beta_Y = -0.2$.

In these simulations there is no genetic component to the moderator (i.e. the MZ correlation always equals the DZ correlation for the moderator). However, genetic effects on the moderator should not have any great impact, unless the genetic effects

Simulated					Estimated				
β_X	β_Y	β_Z	β_M	r	β_X	β_Y	β_Z	β_M	LRT
.	.	.	.	0	0.01	0.00	-0.01	0.00	3.87
				0.5	0.00	0.00	0.00	0.00	3.94
				1	-0.01	0.01	0.00	0.00	4.20
0.2	0.2	0.2	.	0	0.20	0.20	0.21	0.00	211.28
				0.5	0.19	0.21	0.20	0.00	185.22
				1	0.20	0.21	0.20	0.00	168.43
0.2	0.2	0.2	0.2	0	0.20	0.20	0.21	0.20	249.62
				0.5	0.21	0.19	0.21	0.20	217.14
				1	0.22	0.20	0.20	0.20	185.95
0.2	-0.2	.	.	0	0.18	-0.18	0.00	0.00	11.14
				0.5	0.16	-0.17	0.01	-0.01	9.14
				1	0.19	-0.18	0.00	0.00	9.43

Table 4.3: Linear $G \times E$ interaction in twins. The LRT represents the difference in model fit between the ACE and $ACE - XYZ - M$ models (i.e. fixing β_X, β_Y, β_Z and β_M to 0) which is distributed as a χ^2 on 4 degrees of freedom.

are also shared with the trait (this scenario is investigated further below, $G \times E$ in the presence of r_{GE}).

The likelihood ratio test (LRT) column of Table 4.3 represents a 4 degree of freedom test between $ACE - XYZ - M$ and ACE models. As can be seen, the average parameter estimates all fall very close to the simulated values. The LRT is as expected under the null (around 4 for a 4 degree of freedom test). For the combined interactive effects, the LRT values are very high, indicating that a joint test of all moderating effects is very powerful in this case. It appears that moderator variables that are uncorrelated between twins offer slightly more resolving power in $G \times E$ analysis. Finally, note that when the interactive effects are in opposition and almost cancelling each other out, as in the fourth model, the power to detect them jointly is much smaller (although the power to detect them individually would presumably be greater than usual).

4.2.3 Multiple moderator variables

Within this framework, it is possible to incorporate more than one moderator variable along with any interactions between them. As a concrete example, age might moderate genetic effects but only in males. In this case, considering only $A \times M$, two moderator variables, age (M_{Age}, continuous) and sex (M_{Sex}, binary) would have their own interaction coefficients, β_{Age} and β_{Sex}; additionally, an interaction parameter $\beta_{\text{Age}\times\text{Sex}}$ captures any difference in age-moderated genetic effects between sexes. The additive genetic variance component is now

$$(a + \beta_{\text{Age}} M_{\text{Age}} + \beta_{\text{Sex}} M_{\text{Sex}} + \beta_{\text{Age}\times\text{Sex}} M_{\text{Age}} M_{\text{Sex}})^2$$

This kind of extension should probably be limited to cases where prior knowledge or analyses have at least suggested a moderating effect for both variables. The results will be easiest to visualise when one of the moderators is binary, i.e. two plots of variance components as a function of the continuous moderator, one for each level of the binary moderator. A significant interaction parameter for the two moderators means that the slope for a particular variance component will differ between plots.

As an example, a single dataset involving two-variable moderation was simulated. For 500 MZ and 500 DZ twin pairs, continuous (C) and binary (B) moderators were simulated for each individual (only moderation of additive genetic effects is considered in this example). For the '0/1' binary parameter, additive genetic effects were moderated with $\beta_B = 0.5$. The continuous moderator variable had a coefficient of 0.2, but only for individuals scoring '1' on the binary moderator (i.e. $\beta_C = 0$ and $\beta_{B\times C} = 0.2$). Residual components were set at $a = c = e = 1$. The estimates were as follows: $\beta_B = 0.558$, $\beta_C = 0.046$, $\beta_{B\times C} = 0.204$, with $-2LL = 7603.025$. Fixing the interaction moderating parameter, $\beta_{B\times C}$, to 0, the minus twice log-likelihood rose to 7610.854 – a significant difference for 1 degree of freedom.

Summary

This section described a basic $G \times E$ model which allows for one or more continuous or binary moderating variables to have main effects on a trait as well as interacting with any or all of the genetic and environmental latent variables. In general, simulation results suggest that the model will perform well although there may be issues of specificity, e.g. distinguishing between $A \times M$ and $C \times M$.

4.3 Nonlinear $G \times E$ with continuous moderator variables

So far we have assumed that, at the level of effect, all $G \times E$ interactions are linear. In order to more accurately characterise a conceptually interesting class of $G \times E$ models, however, it is necessary to extend the basic model to allow for certain nonlinear interactions. Imagine that at least some level of exposure to an environmental risk is *necessary* to develop disease, whilst high levels of exposure are *sufficient* to cause disease, and that otherwise disease status is influenced by genes. At both extreme low levels of the moderator (nobody has the disease) and extreme high levels (everybody has the disease) there is no genetic variation; at intermediate values of the moderator there is variation due to genes. Although this scenario has been expressed in terms of a binary disease for clarity, similar principles would apply to quantitative traits, as Figure 4.6 illustrates. The leftmost figure illustrates variation (due to genes) only occurring at intermediate values of the moderator, due to the above kind of process. The second figure illustrates the residual variation after the main effect of the moderator has been partialled out in the means model (β_M). This shows the characteristic pattern of genetic variation being attenuated at the extremes of the moderator. The rightmost figure returns to the biometrical model for the hypothetical QTL – there is a nonlinear interaction at the level of effect. This kind of nonlinear interaction can

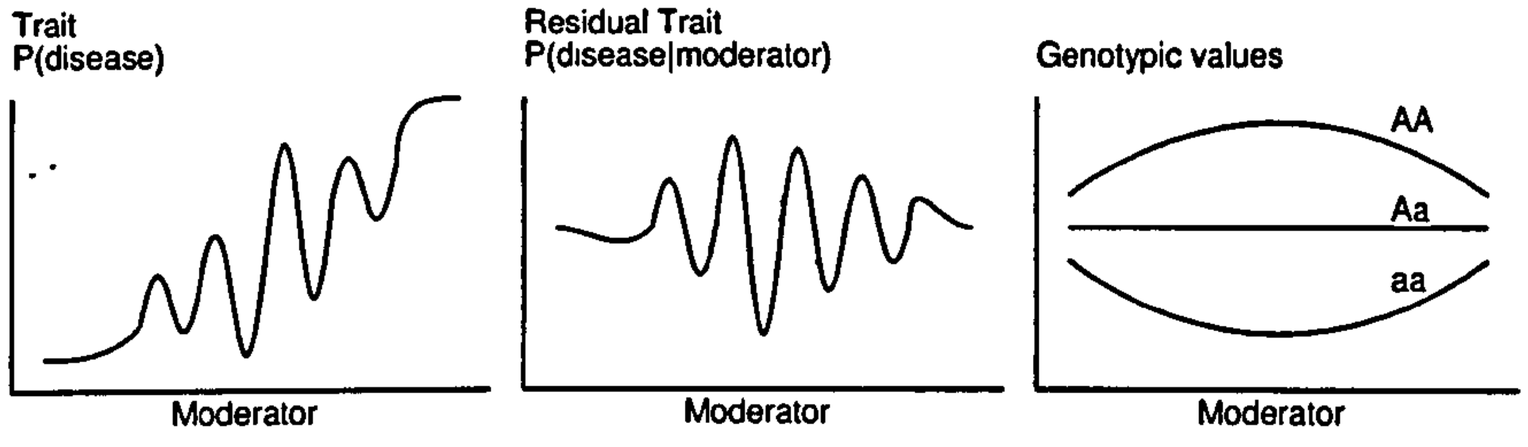


Figure 4.6: Nonlinear interaction and the biometrical model. See text for full description.

be well approximated by a quadratic term being added to the equation describing the additive genetic effect. In other words, the additive genetic path now becomes $(a + \beta_X M + \beta_{X^2} M^2)$. The full model is now $ACE - XYZ - X^2Y^2Z^2 - M$. The variance is

$$Var(T) = (a + \beta_X M + \beta_{X^2} M^2)^2 + (c + \beta_Y M + \beta_{Y^2} M^2)^2 + (e + \beta_Z M + \beta_{Z^2} M^2)^2$$

This kind of model might be of interest in a wide variety of circumstances. For example, it is possible that exposure to combat and post-traumatic stress disorder would follow a similar pattern. Eaves et al (1977) noted that there are many situations in which we might find significant nonlinear trends: "Society may react in a uniform way to extreme deviations on either side of the population mean. This would produce a pattern ... which shows greater environmental variation in the middle of the scale than either end," which would represent nonlinear $E \times E$ in our terminology. The authors continue: "In practice, this kind of interaction is common in psychometric data because of floor and ceiling effects." Alternatively, we may expect genes to operate maximally only in their "average expected environment" (Scarr, 1992) such that genetic variation is attenuated at both environmental extremes.

4.3.1 An example

Twin data were simulated under an AE model such that $r_{MZ} = 0.8$ and $r_{DZ} = 0.4$. On top of these residual components, a moderator variable was simulated with two

properties, (1) a main effect on the trait, such that the phenotypic trait-moderator correlation was $r \approx 0.25$ and (2) a moderating effect such that all genetic effects were exaggerated at intermediate values of the moderator but attenuated at extreme values of the moderator.

This moderating effect is the converse of that in the example linear $G \times E$ simulation, where intermediate values were attenuated and extreme values exaggerated. Note that although the two scenarios display superficial similarity, a nonlinear term is required in this case. For the linear $G \times E$ models, it is clear that (1) the variance components cannot be negative and (2) they are a function of the moderator up to a second-order term. Consequentially, any curve will always be "U-shaped" where the stationary point is also the global minimum. This is not a constraint as such — it follows naturally from assuming a linear $G \times E$ interaction, but it also illustrates the need for the nonlinear models.

Figure 4.7 displays the results for this simulation (leftmost figure), as well as illustrating a further problem with visualising the variance components (other two figures). The best-fitting model is the $AE - X^2 - M$ which precisely recovers the simulated architecture. However, the leftmost figure shows the expected variance components plotted as a function of the moderator, revealing a pattern that only partially corresponds to our intuitions regarding the simulated properties (i.e. exaggerated at intermediate levels, attenuated at extreme levels). That is, the curve describing the A component seems to suggest that, moving in either direction away from the moderator mean, genetic influences decrease and then sharply increase at even more extreme values. Although this is merely an artefact of over-extrapolation and over-fitting, it raises the question of the precise range of the moderator used to visualise the results. In this case, the x-axis corresponds exactly to the observed range of the moderator, which seems a sensible choice (i.e. rather than artificially truncating the moderator distribution). How would one interpret these results in the absence of prior knowledge

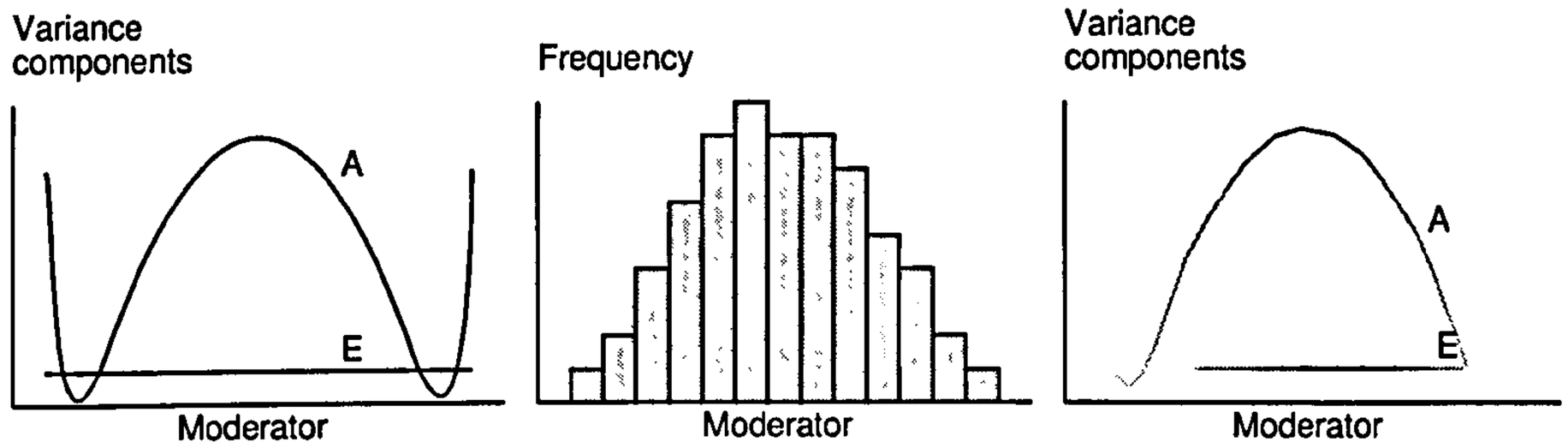


Figure 4.7: Visualisation of variance components for the nonlinear $G \times E$ example.

(i.e. on non-simulated data)?

Consideration of the distribution of the moderator (shown in the middle figure) is useful. As hardly any cases score at those extreme values of the moderator, there is little or no power to estimate the true location of the curve at these points. Although the estimates have over-fit to the data somewhat, there would be very little change to the sample log-likelihood if the curve were drastically altered at these extreme values. A method of visualisation is proposed, and illustrated in the rightmost figure, such that the intensity of the line is directly proportional to the frequency of the moderator within x bins across the distribution. In this way, the visibility of the line is related to the contribution to the sample log-likelihood for that portion of the distribution: if the line is invisible it is because there is little or no power to place it. Although this is not an exact method, it should help to interpret results more clearly, as illustrated in this example, where the curve now approximates the simulated architecture. The Windows `gxe-visualise` program is available for download (`http://statgen.iop.kcl.ac.uk/gxe/`). (For ease of presentation, standard plots will be utilised for the rest of this paper, however.)

4.3.2 Further simulations

A different approach was adopted for this set of simulations, with only 3 replicate datasets simulated under 4 different models in order to allow a closer inspection of

the solutions. All results are presented graphically in Figure 4.8. The four rows of plots represent the four models. The leftmost column of plots represent the true parameter values used to simulate the data. In all cases 500 MZ and 500 DZ twins were simulated, with a moderator variable correlated 0.5 between twins. The middle column represents the variance components estimated in three simulated samples, superimposed upon each other, from the $ACE - XYZ - X^2Y^2Z^2$ model. The third column represents the superimposed $ACE - X - X^2$ model estimates. In all 12 simulations, the $ACE - X - X^2$ model was selected as the best-fitting model, which corresponds to the simulated values chosen.

The four models were chosen to represent different scenarios that involve nonlinear interaction terms. In all cases, the interactions involved only the A additive genetic component; also $c = 1$ and $e = 1.5$. The first model (row a) $a = 2$, $\beta_X = 0$ and $\beta_{X^2} = -0.2$ is similar to the last example. The second model (row b) is similar to the first linear example although involving a quadratic interaction term: $a = 1$, $\beta_X = 0$ and $\beta_{X^2} = 0.2$. The third model (row c) represents a kind of "threshold effect" where only above a certain critical value on the moderator does the genetic variance shoot up (in this case around 1.5 standard deviations above the mean): $a = 0$, $\beta_X = -0.8$ and $\beta_{X^2} = -0.2$. The final model (row d) represents a similar scenario: a threshold type effect for extreme low scorers, but also a linear increasing trend above the mean that plateaus out above 2 SD ($a = 1$, $\beta_X = 1$ and $\beta_{X^2} = -0.2$).

Although not shown on Figure 4.8, other models were fit to the data: in all cases, there was no significant reduction in model fit from dropping the non-genetic interaction terms (i.e. $ACE - XYZ - X^2Y^2Z^2$ versus $ACE - X - X^2$). In contrast, comparing $ACE - XYZ - X^2Y^2Z^2$ and $ACE - YZ - Y^2Z^2$ models, in 11 out of 12 cases the genetic interaction terms could not be dropped. None of the genetic interaction terms could be dropped from the $ACE - X - X^2$ model, however. Using AIC criterion, the $ACE - X - X^2$ model was the best-fit model in all cases, of the

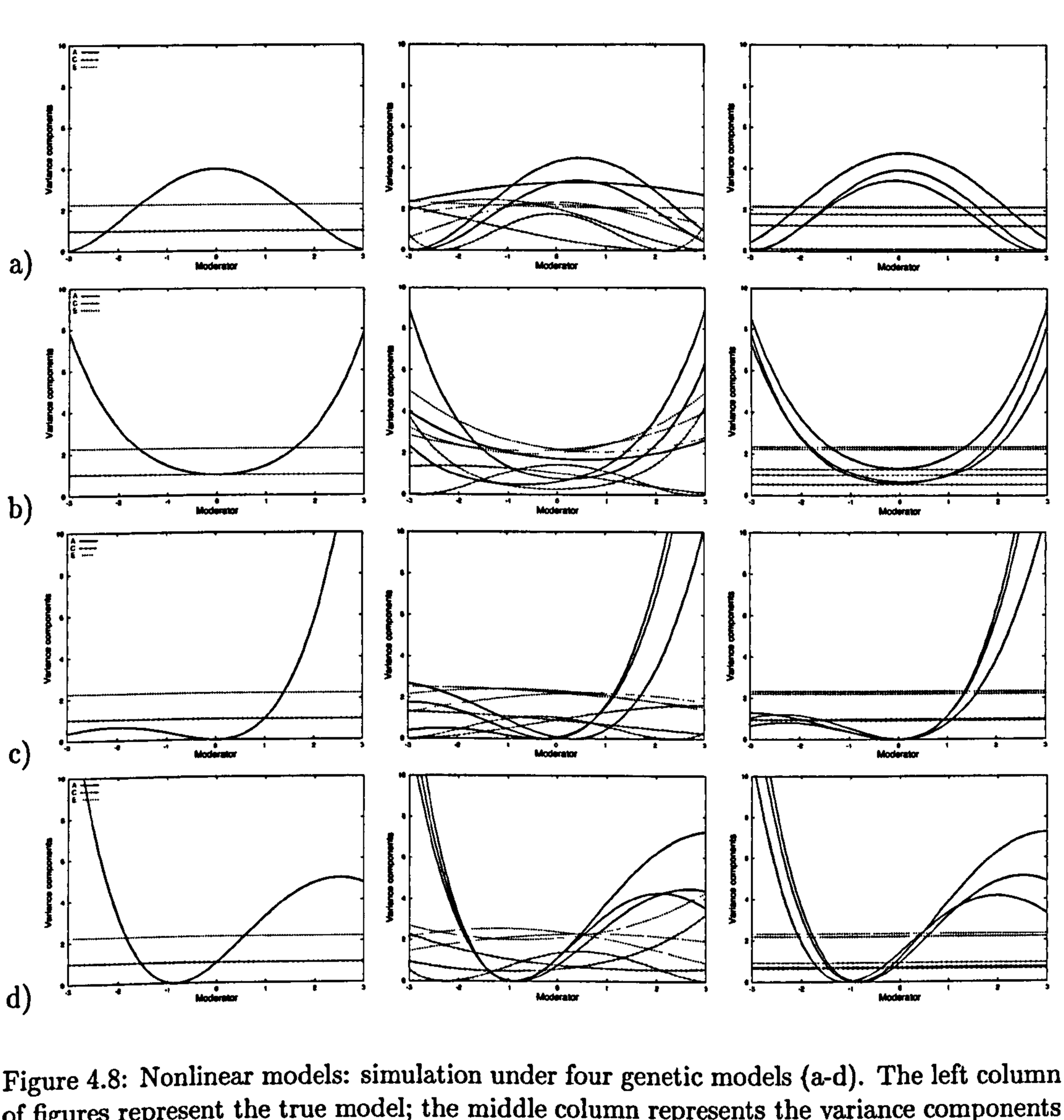


Figure 4.8: Nonlinear models: simulation under four genetic models (a-d). The left column of figures represent the true model; the middle column represents the variance components estimated in three samples of 500 MZ and 500 DZ twins simulated under each model, superimposed on the graph, from the $ACE - XYZ - X^2Y^2Z^2$ model. The third column represents the $ACE - X - X^2$ model estimates from three replicates superimposed – this was the best-fitting model in each case.

models compared: $ACE - XYZ - X^2Y^2Z^2$, $ACE - YZ - Y^2Z^2$, $ACE - X - X^2$, $ACE - X$ and ACE.

As can be seen from Figure 4.8, the parameters are recovered quite well, allowing for sample-to-sample variation. Under the full model the plots are a little messy, but under the $ACE - X - X^2$ model (the best-fitting model in all cases), the simulated structure is recovered very well indeed. The simulated effects are quite large, although nonlinear effects have been found in real, modestly-sized datasets also (unpublished work). Again, it is important to remember whilst looking at the plots, that most of the sample will fall within 1–2 SD above and below the mean, so the models are not quite as extreme as they first seem.

Although generally robust with this sample size, these problems are sensitive to starting values and prone to local minima, as well as being computationally expensive. Care must be taken when fitting these models.

Summary

In order to characterise a large class of potential $G \times E$, in which an effect is attenuated at both high and low extremes of a moderator, a quadratic term was entered in the model. Simulation results suggest that it is possible to discriminate between the nonlinear models and to estimate the interaction coefficients (of which there are up to 6) quite well using only moderately large sample sizes.

4.4 Gene–environment correlation

Potential moderators will typically be correlated with the trait – in the absence of strong *a priori* reasons, it is likely to be the phenotypic association that flagged up the variable as a potential moderator in the first case. Although this correlation may be due to trait-mediating effects of the moderator, it may alternatively be due to

other shared causes, which includes the possibility of shared genes. It is well known that many "environmental" variables demonstrate substantial heritable components (Plomin et al., 2001). Many environmental variables may in fact be *correlated* with the genetic effects on the trait (r_{GE}) rather than *modifying* the genetic effects on the trait ($G \times E$). As mentioned earlier, r_{GE} can appear as $G \times E$ in typical analyses of $G \times E$. However, the current approach can be easily extended to model $G \times E$ in the presence of r_{GE}.

Entering the moderator in the means model to allow for a main effect will effectively remove from the covariance model any genetic effects that are shared between trait and moderator. That is, r_{GE} will appear as a main effect, β_M, due to the moderator acting as a proxy measure for the additive genetic effects on the trait. Any interactions detected will not be due to r_{GE}, but rather will be interactions between the moderator and variance components specific to the trait. In this way, evidence for $G \times E$ will never reflect a "false-positive" claim for interaction. However, this approach will also fail to detect $G \times E$ interaction where the moderated genetic component is common between trait and moderator, i.e. $G \times E$ in the presence of r_{GE}.

Table 4.4 illustrates the application of the basic $G \times E$ model presented in this Chapter so far in the presence of a genetic correlation between moderator and trait. Again, 500 MZ and 500 DZ twins were simulated 50 times under each condition; $a = c = e = 1$. Note that the estimate of β_M is inflated due to the shared genetic effects. More importantly, the tests of $G \times E$ not allowing for any main effect (the last column) show inflated test statistics when there is no interaction but there is a gene–environment correlation. That is, the third and sixth rows have average values of 4.01 and 11.06 for this test, both of which are greater than the critical value for this 1 degree of freedom test. Note however that the test of an interaction that allows for a main effect (second to last column) does not show such an effect.

Simulated			Estimated					Likelihood ratio tests		
								$ACE-X-M$	$ACE-X-M$	$ACE-X$
r_{GE}	β_X	β_M	a	c	e	β_X	β_M	$ACE-X$	$ACE-M$	ACE
0	.	.	0.99	0.99	0.99	-0.01	0.00	0.83	1.03	1.03
0.5	.	.	0.95	0.98	1.01	0.00	0.14	37.86	1.04	1.07
1	.	.	0.66	1.05	1.04	0.00	0.28	155.51	1.76	4.01
0	.	0.2	0.99	1.01	1.00	0.00	0.20	79.66	0.99	0.95
0.5	.	0.2	0.93	1.02	1.00	0.00	0.34	215.67	1.06	2.07
1	.	0.2	0.65	1.06	1.04	0.00	0.48	406.74	1.36	11.06
0	0.2	.	0.99	1.00	1.00	0.21	0.00	0.97	49.60	49.73
0.5	0.2	.	0.94	1.00	1.00	0.20	0.12	25.72	46.54	55.07
1	0.2	.	0.68	1.05	1.02	0.25	0.25	100.40	46.19	88.90
0	0.2	0.2	1.01	0.98	0.99	0.21	0.20	74.03	52.61	56.26
0.5	0.2	0.2	0.94	1.00	1.00	0.20	0.31	167.17	44.99	77.03
1	0.2	0.2	0.70	1.04	1.02	0.25	0.45	308.27	43.88	146.27

Table 4.4: Performance of the basic $G \times E$ model in the presence of r_{GE} (i.e. the moderator M has shared genetic influence with the trait, measured by the genetic correlation, r_{GE}). The table presents parameter estimates for the $ACE-X-M$ model and three likelihood ratio tests: of a main effect in the presence of an interaction; of interaction in the presence of a main effect; of interaction not allowing for any main effect.

4.4.1 $G \times E$ in the presence of r_{GE}

In the previous simulations, the interactive effect was simulated for the genetic effects specific to the trait. If in fact the interaction was with only the genetic effect shared with the moderator, the above model would have failed to detect it. As mentioned, this is because these effects have already been partialled out in the means model. The current model can be re-formulated as a bivariate model of both trait and moderator in order to detect these effects of $G \times E$ in the presence of r_{GE}, however. Figure 4.9 shows the partial path diagram for one twin to illustrate this approach. Here the moderator features twice in the model – as a dependent variable to be modelled as well as a moderator variable to define the paths to the trait. The main effect in the means model has been replaced with a path indicating shared genetic effects. The trait is now influenced by two sources of genetic influence: that which is shared with the moderator, and that which is not (common and unique paths a_C and a_U). Each path can interact with the moderator, represented by the coefficients β_{X_C} and β_{X_U}. The C and E components (not shown on the path diagram) follow the standard bivariate Cholesky parameterisation, but without moderation. From these parameters, r_{GE}

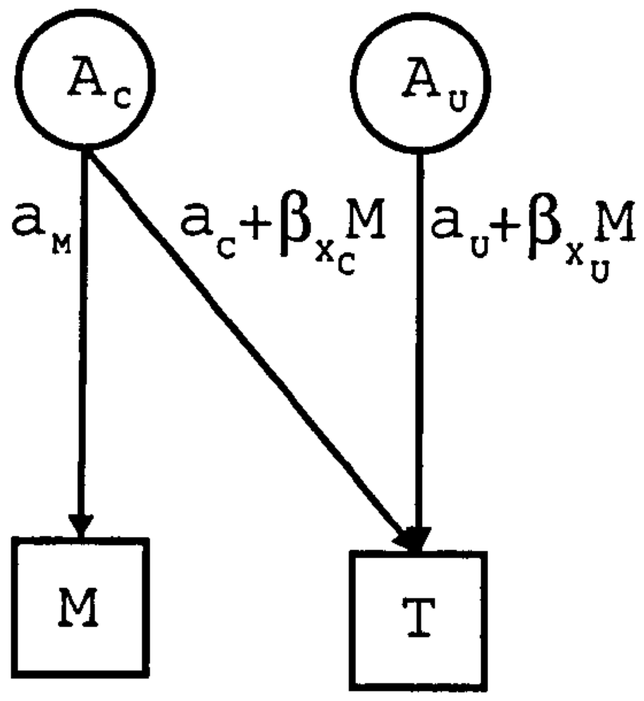


Figure 4.9: Extended $G \times E$ model to allow for gene–environment correlation.

can be calculated as $a_M a_C / \left(a_M \sqrt{a_C^2 + a_U^2}\right)$ if there is no $G \times E$ but otherwise gives r_{GE} at the mean value of the moderator (assuming a zero-centred moderator). In the presence of $G \times E$, r_{GE} must be calculated conditional on M as it will also vary as a function of the moderator

$$r_{GE}|M = \frac{a_M(a_C + \beta_{X_C}M)}{a_M\sqrt{(a_C + \beta_{X_C}M)^2 + (a_U + \beta_{X_U}M)^2}}$$

whilst calculating the average r_{GE} for the sample involves integrating over the distribution of the moderator.

The following set of simulations illustrates this model's ability to distinguish between the two types of interaction, whether or not there is a genetic correlation. Results are shown in Table 4.5. The conditions (each consisting of 25 replicate samples of 500 MZ and 500 DZ twins) varied the genetic correlation between moderator and trait and the presence or absence of moderating effects on the common and unique genetic paths, as described above. The genetic correlations correspond to the correlation of unmoderated effects only. That is, $r_{GE} = 0$ corresponds to $a_M = 1$, $a_C = 0$ and $a_U = 1$; $r_{GE} = 0.5$ corresponds to $a_M = 1$, $a_C = \sqrt{0.25}$ and $a_U = \sqrt{0.75}$; $r_{GE} = 1$ corresponds to $a_M = 1$, $a_C = 1$ and $a_U = 0$. Shared and nonshared environmental effects were simulated for each component to have a variance of 1 and be uncorrelated

Simulated			Estimated					Likelihood ratio tests		
								$ACE - X_C X_U$	$ACE - X_C X_U$	$ACE - X_C X_U$
r_{GE}	β_{X_C}	β_{X_U}	a_M	a_C	a_U	β_{X_C}	β_{X_U}	ACE	$ACE - X_U$	$ACE - X_C$
0	.	.	0.99	-0.01	0.97	0.01	0.01	2.83	1.77	1.03
0.5	.	.	0.98	0.48	0.78	0.01	0.00	1.72	0.92	0.83
1	.	.	1.04	0.98	0.26	0.00	0.03	1.92	0.87	1.04
0	0.2	.	1.01	-0.01	0.94	0.20	0.01	59.74	57.72	1.27
0.5	0.2	.	1.04	0.49	0.84	0.19	0.00	63.82	53.67	1.00
1	0.2	.	1.03	0.94	0.29	0.20	0.03	86.77	53.99	0.70
0	.	0.2	1.02	0.00	0.99	0.00	0.19	46.28	0.65	26.81
0.5	.	0.2	1.03	0.45	0.88	0.01	0.20	41.68	1.04	22.80
1	.	0.2	1.00	0.94	0.26	0.00	0.18	8.43	1.14	6.85
0	0.2	0.2	1.05	-0.07	0.97	0.19	0.22	99.28	44.28	19.64
0.5	0.2	0.2	1.02	0.48	0.88	0.19	0.20	123.36	45.12	8.82
1	0.2	0.2	1.00	0.93	0.20	0.19	0.17	86.29	40.42	5.88
0	0.2	-0.2	1.04	-0.01	0.99	0.18	-0.20	94.53	40.40	15.33
0.5	0.2	-0.2	1.02	0.44	0.88	0.21	-0.19	80.46	49.79	15.49
1	0.2	-0.2	0.99	1.00	0.18	0.20	-0.20	86.30	47.22	4.48

Table 4.5: Performance of the extended $G \times E$ model in the presence of r_{GE}. The table presents parameter estimates for the $ACE - X_C X_U$ model and three likelihood ratio tests: of moderation for both common and unique genetic effects; of moderation for unique effects only; of moderation for common effects only.

between trait and moderator (i.e. $c_M = e_M = 1$, $c_C = e_C = 0$ and $c_U = e_U = 1$). The 15 conditions are arranged in five blocks: (1) no moderation (2) moderation of common path (3) moderation of unique path (4) moderation of common and unique path, similar effects (5) moderation of common and unique paths, opposing effects. Four models were analysed: $ACE - X_C X_U$, $ACE - X_C$, $ACE - X_U$ and ACE. Three likelihood ratio test statistics were constructed (final three columns in Table 4.5): in order of the columns (1) moderation for both common and unique genetic effects (2) moderation for unique effects only (3) moderation for common effects only. The parameter estimates under the full $ACE - X_C X_U$ model are also shown in the Table.

Under the null (first three rows), the models perform as expected: the β coefficients are all near zero, and the tests of moderation show average χ^2 values close to their expected values under no moderation. The average unmoderated genetic parameter values are close to their simulated values, with the exception of a_U when $r_{GE} = 1$, which is simulated at 0, but has an average estimated value of 0.26. This artefact, which also exists in the other $r_{GE} = 1$ conditions, is explained further below.

The second three conditions simulate a moderating effect of shared genetic influence between the trait and moderator. This effect is recovered well and detected, no matter what the background genetic correlation between trait and moderator. The specific tests of β_{X_C} (the second column of likelihood ratio tests in the Table) show highly significant average values, whereas the β_{X_U} parameters average near zero. Power seems to increase as r_{GE} increases. The third three conditions simulate a moderating effect of genetic influence specific to the trait. Again, the parameters are recovered well, although power to detect β_{X_U} drops off as r_{GE} increases. The final six rows of Table 4.5 show that the model works when both β_{X_C} and β_{X_U} are nonzero.

When $r_{GE} = 1$ some subtle properties of the model emerge – they are worth considering in further detail. The contribution to the variance of total unique genetic effects is $(a_U + \beta_{X_U} M)^2 = a_U^2 + 2a_U\beta_{X_U} M + \beta_{X_U}^2 M^2$. As mentioned above, the power to detect β_{X_U} decreases with increasing r_{GE}, because when $r_{GE} = 1$ then $a_U \rightarrow 0$ and so $2a_U\beta_{X_U} M \rightarrow 0$ which reduces the impact of β_{X_U} on the variance by cancelling this cross-product term. A similar logic applies to the relationship between a_C and r_{GE}.

Additionally, as a_U and therefore $2a_U\beta_{X_U} M$ approach 0, then β_{X_U} only makes squared contributions to the variance. Therefore, when $r_{GE} = 1$, the estimate of a_U is likely to be near zero, which reduces the power to identify the *sign* of β_{X_U} although the *absolute value* can still be identified. Although the contribution to the variance will be the same (and so this is not an issue for the analysis of real data), taking the average of the unsquared parameter in repeated simulation would lead to an apparent bias in parameter estimate for β_{X_U} when $r_{GE} = 1$. The average values for β_{X_U} (not shown in the Table) were in fact -0.03, 0.02 and -0.12, in the 9^{th}, 12^{th} and 15^{th} rows, respectively: making the signs all positive (or all negative in row 15) produces the unbiased average parameter estimates as shown in the Table (0.18, 0.17 and -0.20).

As noted above, there is also an apparent bias in the estimates of a_U when $r_{GE} = 1$. This parameter has a large standard error, and it can be fixed to 0 when $r_{GE} = 1$ with

no significant reduction in fit, on average. Again, under certain conditions the sign of a_U is not identified; however, optimisation favours the positive values, probably due to a positive starting value being specified. This apparent bias is therefore not important in real analysis, it is only a consequence of taking averages.

The current model of $G \times E$ in the presence of r_{GE} could be extended in a number of ways. For example, a third variable that is a potential index of genetic sensitivity to an environmental factor can be incorporated, to produce models similar to recent Markov Chain Monte Carlo methods which handle $G \times E$ in the presence of r_{GE} (Eaves and Erkanli, 2002).

Summary

In the basic model, any genetic effects that are shared between the trait and the moderator will be modelled as main effects of the moderator. An extension to the basic model explicitly models shared genetic effects, as well as any interactions between these effects and the moderator, allowing for the analysis of $G \times E$ in the presence of r_{GE}.

4.5 Qualitative $G \times E$ with continuous moderator variables.

All the previous models of $G \times E$ have implicitly addressed *scalar*, or *quantitative*, moderation, meaning that the *magnitude* of polygenic effect has varied as a function of the moderator. However, it is also possible that *different* polygenic effects operate at different points along the moderating continuum. The same distinction is found in 'sex-limitation' models, where males and females may have different magnitudes of genetic (or environmental) effects but may also differ in which genes operate in males and females. Evidence for the 'different genes' hypothesis is reduced covariance

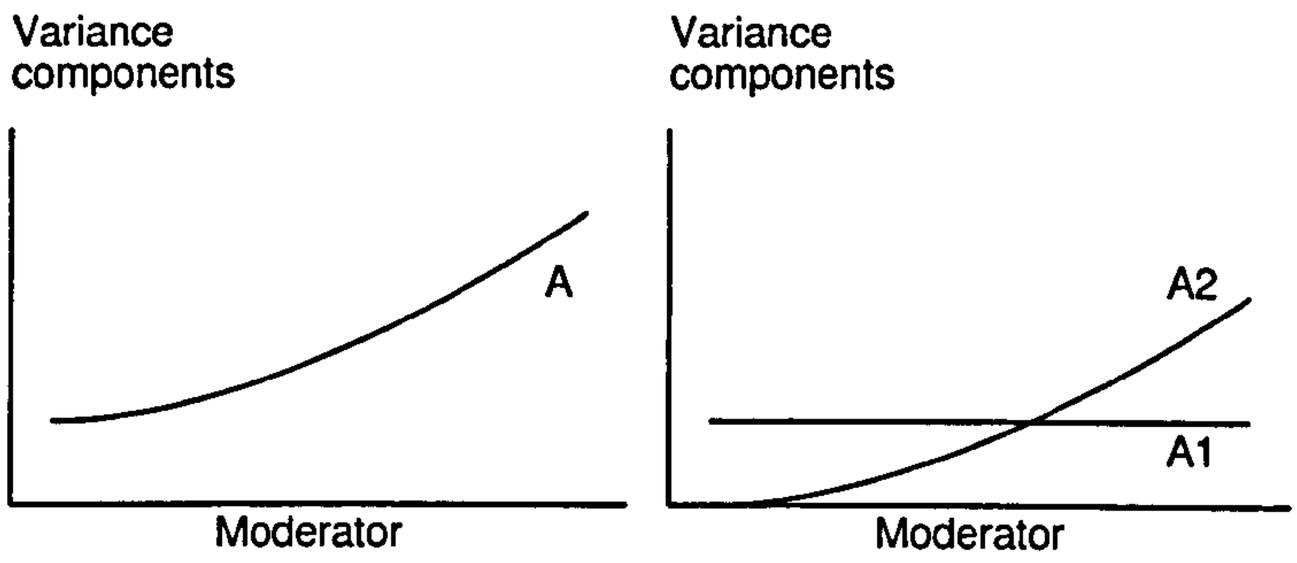


Figure 4.10: Schematic illustrating scalar (left figure) and qualitative (right figure) $G \times E$. See text for further explanation.

between twins discordant for the moderating variable, i.e. sex.

To allow for qualitative $G \times E$ with continuous moderator variables, we adopt the most simple biological model: that there are two independent sets of polygenes, $A1$ and $A2$, which show different profiles of scalar interaction with the moderator. Figure 4.10 illustrates this concept. The left panel depicts a standard moderated variance component, which is consistent with at least some genetic effects being amplified at higher values of the moderator. This curve could also have come about as a result of different genes operating at higher levels of the moderator, however, as shown in the right panel. Here we see that the $A1$ set of polygenes is not moderated, whereas the $A2$ set only have an effect at high levels of the moderator. In this way, individuals high on the moderator have a different profile of genes operating compared to individuals low on the moderator (not just greater or lesser effects of the same genes).

It is worth drawing a distinction between qualitative interaction and r_{GE}. Qualitative interaction implies that different loci have an effect depending on the value of the moderator. Gene-environment correlation implies that certain alleles of certain loci are present depending on the value of the moderator. In the latter case, an association between an individual's genetic loading and the moderator ensues, which has to be explicitly modelled. This is not the case for qualitative interaction however.

Noting that a model with both sets of polygenes showing scalar interaction is not

identified, the expected additive genetic variance for twin i is now $a_1^2 + (a_2 + \beta_{X_2} M_i)^2$; the additive genetic component of the MZ covariance is $a_1^2 + (a_2 + \beta_{X_2} M_1)(a_2 + \beta_{X_2} M_2)$; the additive genetic component of the DZ covariance is $a_1^2/2 + (a_2 + \beta_{X_2} M_1)(a_2 + \beta_{X_2} M_2)/2$. The formulation implies that the effective coefficients of genetic relatedness will be attenuated for twin pairs discordant on the moderator under qualitative interaction. That is, the effective coefficients (normally 1 and 0.5 for MZ and DZ pairs respectively) are for MZ pairs

$$\alpha_{MZ} = \frac{a_1^2 + (a_2 + \beta_{X_2} M_1)(a_2 + \beta_{X_2} M_2)}{\sqrt{a_1^2 + (a_2 + \beta_{X_2} M_1)^2}\sqrt{a_1^2 + (a_2 + \beta_{X_2} M_2)^2}}$$

and for DZ twins

$$\alpha_{DZ} = \frac{a_1^2/2 + (a_2 + \beta_{X_2} M_1)(a_2 + \beta_{X_2} M_2)/2}{\sqrt{a_1^2 + (a_2 + \beta_{X_2} M_1)^2}\sqrt{a_1^2 + (a_2 + \beta_{X_2} M_2)^2}}$$

This model will be referred to as the $A_1 A_2 CE - X_2$ model (assuming that shared and nonshared environmental components are also included). It can be seen that when $M_1 = M_2$ then $\alpha_{MZ} = 1$ and $\alpha_{DZ} = 0.5$ for any values of a_1, a_2 and β_{X_2}. Figure 4.11 shows the attenuation for MZ and DZ pairs as a function of M_1 and M_2. Note that the exact shape of this surface will depend on a_1, a_2 and β_{X_2} and can go negative under certain conditions. Clearly, this model is not applicable for obligatorily shared moderators.

Initial simulation results suggest poor power to discriminate between scalar and qualitative $G \times E$, however. Table 4.6 shows the results fitting the $ACE - X$ (scalar) and $A_1 A_2 CE - X_2$ (qualitative) models to six example datasets simulated under either scalar $G \times E$ (first three rows) or qualitative $G \times E$ (second three rows). The likelihood ratio test statistic (LRT column) is the χ_1^2 increase in fit from qualitative to scalar models. Simulating under a population value of $a_1 = 0$ implies scalar interaction (i.e. there is only one set of polygenes). Also, note that $c = e = 1$ and that 1000

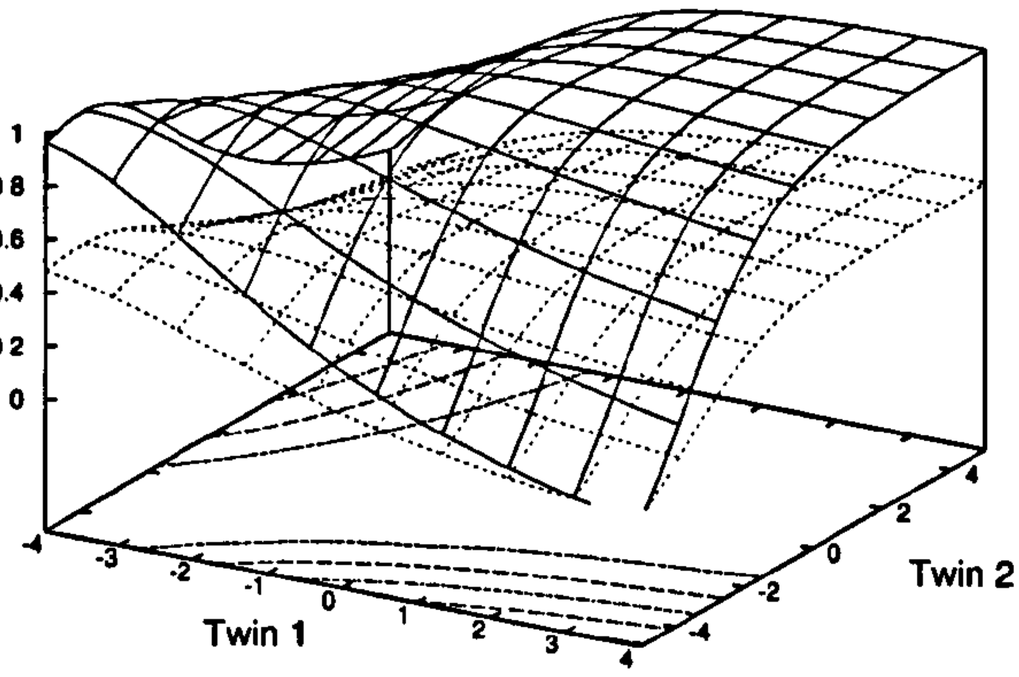


Figure 4.11: Plot of α_{MZ} (solid grid) and α_{DZ} (dotted grid) for $a_1 = a_2 = 1$ and $\beta_{X_2} = 0.3$. Along the diagonal $M_1 = M_2$, $\alpha_{MZ} = 1$ and $\alpha_{DZ} = 0.5$.

Simulated			$A_1A_2CE - X_2$			$ACE - X$		LRT	
a_1	a_2	β_{X_2}	a_1	a_2	β_{X_2}	a	β_X	χ^2	p
0	1	0.2	0.21	1.04	0.20	1.06	0.19	0.00	0.95
			0.00	1.10	0.17	1.10	0.17	0.00	1.00
			0.00	0.83	0.26	0.83	0.26	0.00	1.00
1	0.5	0.25	1.08	0.44	0.22	1.16	0.09	2.56	0.11
			1.14	0.45	0.19	1.22	0.07	1.51	0.22
			0.95	0.39	0.28	1.02	0.11	5.95	0.01

Table 4.6: Scalar and qualitative $G \times E$: results from six simulated example datasets.

pairs of each zygosity were simulated (twice the usual sample size). As can be seen, the qualitative model is correctly rejected in all three scalar cases (first three rows). However, there is at best only very weak evidence to support the qualitative model in the second three datasets, with only 1 of the three being significant at the 5% significance level. More extensive simulation work is required to properly evaluate the power of this test under a range of conditions.

Summary

An interaction may involve the same genes having different effects (scalar interaction) or different genes operating (qualitative interaction) at different levels of the moderator. A simple model of qualitative interaction for continuous variables was presented, although power to discriminate between scalar and qualitative interaction empirically looks likely to be low.

4.6 Other distributional factors influencing $G \times E$ analysis

4.6.1 Mismatching continuous and binary moderators

Although many moderator variables may act continuously, it is also entirely reasonable that some moderators act in a more discrete manner, even if they can be measured on a continuous scale. In this section we consider the impact of 'misclassifying' a moderator variable: either falsely dichotomising what is actually a continuous moderator or using a continuous measure when the moderating effect is really a threshold effect (e.g. only the top 10% individuals show an increased genetic effect).

Samples were simulated under two kinds of model: continuous or binary moderation. For the continuous case, $\beta_X = 0.2$; for the binary case, the continuous moderator was transformed to a binary scale, with individuals more than 1.28 standard deviations above the mean scored "1", all others scored "0" (corresponding to a 9:1 ratio of "0":"1"), with $\beta_X = 0.8$ (no direct comparison can be made between the magnitude of interaction in the continuous and binary cases in terms of β_X alone however).

Similarly, analysis adopted either a continuous or binary approach towards the moderator. The correctly classified scenarios are therefore when the data were generated using a continuous moderator and also analysed using a continuous moderator;

		Model											
		Continuous					Binary						
		Linear		Nonlinear			10%		5%		25%		
Data	r	β_X	LRT	β_X	β_{X^2}	LRT	β_X	LRT	β_X	LRT	β_X	LRT
Continuous	0	0.20	22.02	0.20	0.00	23.36	0.36	8.17	0.41	6.88	0.33	12.58
	0.5	0.20	18.05	0.20	0.01	19.16	0.36	7.72	0.40	5.67	0.32	10.82
	1	0.22	17.04	0.23	-0.02	18.00	0.37	6.53	0.39	4.74	0.35	10.03
Binary (10%)	0	0.17	16.12	0.15	0.10	27.40	0.79	41.30	0.71	17.95	0.37	16.12
	0.5	0.24	18.57	0.18	0.10	28.80	0.84	42.84	0.73	18.62	0.39	16.36
	1	0.20	12.63	0.15	0.10	19.88	0.81	28.77	0.73	13.30	0.35	10.71

Table 4.7: Continuous and binary moderators: effects of misspecifying moderator type. The LRT represents the difference in model fit between the ACE and $ACE - X$ (or $ACE - X - X^2$) models. All LRT are distributed as a χ^2 on 1 degree of freedom, except for the nonlinear test which is on 2 degrees of freedom.

also, when the data were generated using a binary moderator and analysed using the same binary moderator. The misclassified scenarios are when the data were generated using a continuous moderator, which was subsequently dichotomised for analysis; also, when the data were generated using a binary variable but the analysis used the underlying continuous 'liability' instead.

In addition, some further analytic conditions were considered. A binary moderating effect could rightly be described as 'nonlinear' in terms of the underlying continuous dimension – the nonlinear model was therefore included when analysing a continuous moderator to see how well a quadratic function performs at approximating the 'step function' of a threshold effect. Finally, although it is common for experimenters to dichotomise continuous or semi-continuous variables in analysis (e.g. taking the top 5% of high scorers on the moderator) the chosen threshold may not correspond to the actual threshold (e.g. actually the top 10% show a moderated effect). Two further analysis conditions were included to represent this kind of misclassification of binary variables: the true threshold was 10% and the classification was either too harsh (5%) or too inclusive (25%). In all cases, $a = c = e = 1$ for 500 MZ and 500 DZ twin pairs; 50 replicates were generated under each condition.

Table 4.7 shows that, as expected, a continuous analysis model works much better when the moderation is truly continuous; likewise, a correctly specified binary moder-

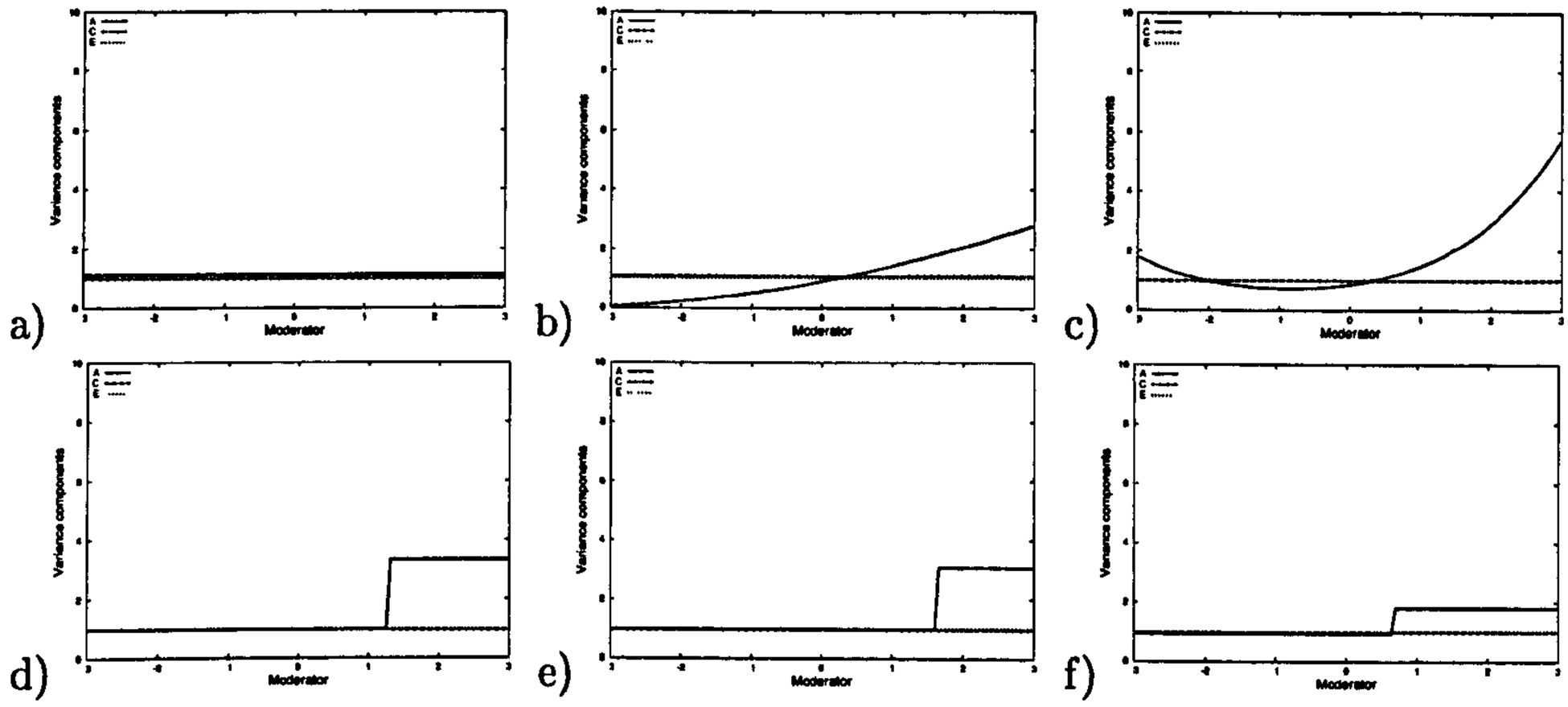


Figure 4.12: Binary moderators and continuous approximations, for data simulated with a binary moderating effect (10% threshold): (a) ACE model (b) $ACE - X$ model with continuous moderator (c) $ACE - X - X^2$ model with continuous moderator (d) $ACE - X$ model with binary moderator (10%) (e) $ACE - X$ model with binary moderator (5%) (f) $ACE - X$ model with binary moderator (25%).

ator in analysis performs best when the moderation is truly binary. For continuously-moderated data, the average test statistic under the continuous analysis models is typically at least double the binary analysis models. As expected, allowing for a non-linear continuous effect adds nothing. For binary-moderated data, the 10% binary model in analysis works best. However, the nonlinear model seems to offer a good approximation, capturing around three-quarters of the available information. Furthermore, when the binary analysis model is misspecified (i.e. the dichotomy is either too harsh or too inclusive), then performance is worse than the nonlinear model and equal to the linear continuous model.

From these results it seems to be a good strategy to adopt continuous moderators whenever available, allowing for nonlinear moderation to model any threshold effects. Figure 4.12 shows the average estimated variance components as a function of the moderator under different analysis models for the case of a binary moderating effect in the data, based on a liability with a sib correlation of 0.5.

4.6.2 Non-normal trait distributions

The current method relies on often relatively subtle differences in the variance, MZ covariance and DZ covariance across the range of the moderator variable to infer the presence of any interactive effect. Whilst it would be expected that deviations from multivariate normality may obscure these subtle effects, it is also possible that certain forms of measurement bias and error could lead to spurious evidence for $G \times E$.

Many behavioural measurements have skewed, or J-shaped, distributions. For example, on a six-point symptom scale, the majority of individuals might score 0 or 1, whilst only a handful of individuals score above 4. If such a measure does in fact represent of normally-distributed liability, then the low end of the scale distribution is less informative than the high end. If a second variable correlates with the trait, then the second variable will also correlate with the 'informativeness' of the first measure. This would be detected as an interactive effect. For example, the second variable would predict that twins with similar low scores on the liability are more likely to have identical scores on the measurement than twins with similar high scores on the liability. This would be an example of heteroscedasticity.

A set of simulations investigated this effect. In all cases, $a = c = e = 1$ for 500 MZ and 500 DZ twin pairs. A continuous covariate was simulated with a sibling correlation of 0.5. Three conditions were assessed: (1) no moderation and no main effect, $\beta_X = \beta_M = 0$ (2) a main effect only, $\beta_X = 0$, $\beta_M = 0.5$ and (3) a true moderating effect and a main effect, $\beta_X = 0.2$, $\beta_M = 0.5$. Twenty-five replicate datasets were simulated under each of the three models. In analysis, two models were fit to the data: $ACE - XYZ - M$ and $ACE - M$, the difference in fit between which provides a 3 degree of freedom test of *any* moderating effect. Each replicate dataset was subjected to two transformation schemes, illustrated in Figure 4.13. Transformation 1 simply bins the continuous trait score into a less informative 15-point scale; transformation 2 bins the scores more severely and introduces a skew into the distribution.

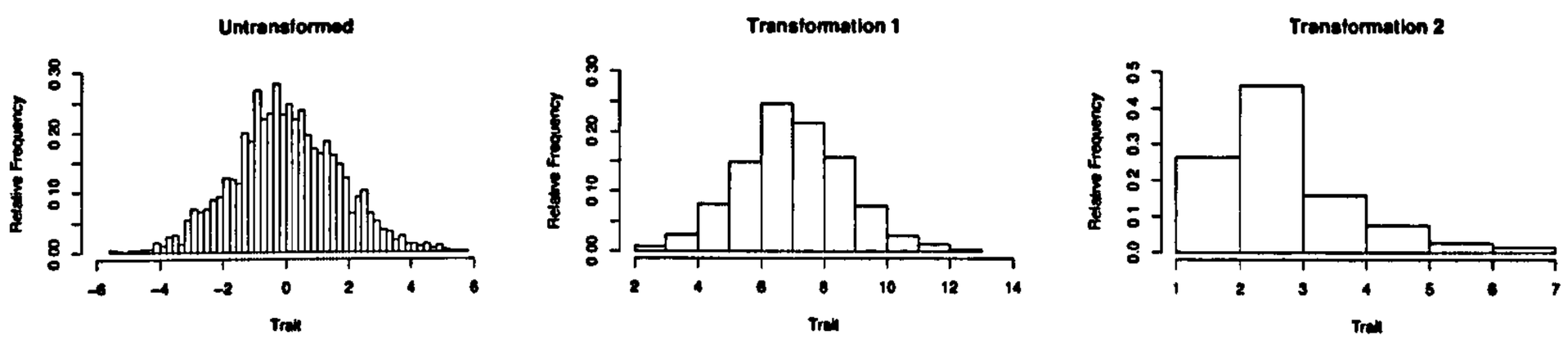


Figure 4.13: Example of data simulated and then transformed to investigate tests of moderation in skewed distributions. The first transformation bins the datapoints into 15 bins; the second transformation is more severe and introduces a skew in the data.

Simulated		Estimated				
β_M	β_X	β_M	β_X	β_Y	β_Z	LRT
Untransformed						
.	.	0.00	0.02	-0.02	-0.02	2.88
0.5	.	0.51	-0.02	0.01	0.01	3.67
0.5	0.2	0.50	0.16	0.03	0.01	19.16
Transformation 1						
.	.	0.00	0.03	-0.02	-0.02	2.74
0.5	.	0.50	-0.03	0.02	0.00	3.65
0.5	0.2	0.50	0.16	0.03	0.02	18.32
Transformation 2						
.	.	0.00	0.00	-0.01	0.00	3.60
0.5	.	0.28	0.06	0.09	0.08	60.50
0.5	0.2	0.30	0.17	0.10	0.09	126.69

Table 4.8: Tests of moderation under skewed trait distributions. The LRT column is the 3 degree of freedom likelihood ratio tests statistic for $\beta_X = \beta_Y = \beta_Z = 0$.

Table 4.8 gives the results for this set of simulations. We would not expect the likelihood ratio test of $ACE - XYZ - M$ against $ACE - M$ to be significant for scenario (1) or (2), whereas it should be significant for (3) due to the simulated interaction. This holds for the untransformed data and under the first transformation (the critical value for a χ^2 statistic with 3 df at the $\alpha = 0.05$ level is 7.815), but not under the second transformation scheme: the difference in fit is 60.5 which has a p-value of 4.6×10^{-13} for condition (2) where no interaction is actually simulated. In other words, the transformation scheme has induced evidence for some kind of moderating effect.

This effect may seem to be a cause of concern, given high prevalence of such measurement scales. Inspection of the moderating coefficients should reveal a predictable signature however, $\beta_X \approx \beta_Y \approx \beta_Z$ when $a \approx c \approx e$. Plotting the expected variance components will reveal only a gentle trend for all variance components to be attenuated similarly at low levels of the moderator. Whilst possible as a real model, researchers should be cautious in their conclusions, especially in the presence of heteroscedasticity. A scenario when β_X, β_Y and β_Z are all significant in a similar direction is also consistent with what might be called "phenotypic interaction" between the trait and the moderator, or "$P \times E$". In this case, the moderator doesn't interact with any component of variance specifically; rather, it increases variation in the entire trait, at what can be thought of as a 'later stage' in the trait's aetiology.

Summary

The distributions of the moderator (binary versus continuous) and of the trait (normal versus non-normal) were investigated in this section. It appears that, under a nonlinear model, using a continuously measured moderator works well even if the actual moderation operates as a binary threshold effect. It was also shown how certain types of skewed trait distributions might generate spurious evidence for interaction.

4.7 Discussion

As long as an individual's genetic makeup is represented by a single, latent '*A*', then the possibilities of gene–environment interaction will approach an unavoidable limit. In the future, the analysis of multiple measured genotypes interacting with multiple measured environmental factors will be necessary, in order to refine the broad brush-strokes we currently use to characterise the quantitative genetics of complex human traits. Nonetheless, twin analysis of gene–environment interaction using continuous moderator variables should still offer some interesting insights into the aetiology of many complex traits, although several issues not yet covered may emerge in the application of these models to real data.

The simulations presented in this Chapter generate data that is 'cleaner' than we might expect in practice. Although this is typically the case with all simulation studies, the present models rely on relatively subtle phenomena and so the extent to which systematic and stochastic biases generate misleading results has not been fully addressed. Most simulations were conducted using a moderately large sample of 500 MZ twins and 500 DZ twins: the behaviour of the models in smaller and larger samples is of interest, also.

In addition to its cleanliness, a simulated dataset comes with the knowledge of the true model, which inevitably guides analysis. In practice, for a specific dataset it might not be obvious how best to approach the various inter-related questions that can be asked: binary versus continuous moderation, linear versus nonlinear effects, interactions versus main effects versus correlations, scalar versus qualitative interactions, multiple moderators, etc. It might therefore be useful to develop a 'protocol', by which different models are sensibly and systematically evaluated and compared.

Although standard bivariate models explain the relationship between any two traits in terms of shared or direct causation, the kind of relationship involved in $G \times E$ might also be plausible. In other words, it is not necessary that the E component

of $G \times E$ actually be "environmental" in any traditional sense of the word. What constitutes an environment from the gene's point of view is quite different from an individual's point of view. For example, the internal biochemical state of the body in which a gene finds itself can sensibly be called it's environment. For appropriate traits, it might therefore be worth considering the above interaction models along with the standard bivariate ones. Consider a fictitious example involving anorexia and neurotic symptoms. Say being anorexic has various consequences including chronic low body weight. Low body weight may in turn lead to genes being switched on or off, some of which might operate to increase or decrease the chance of neurotic symptoms. Therefore, there will be an increase in the genetic variance of neuroticism, as a consequence of an anorexia-related state switching on genes. This scenario is distinct from having a set of genes that operate jointly on anorexia and neuroticism (i.e. a genetic correlation); it is distinct from direct causation between anorexia and neuroticism, in that although being anorexic leads to an increased risk of being neurotic, this is only expressed in genetically-predisposed individuals. As such, this dynamic fits within the same analytic framework as the $G \times E$ models considered so far: in this example, a $G_{\text{Neuroticism}} \times E_{\text{Anorexia (}\rightarrow\text{body weight)}}$ interaction. Such an effect might be called a "Gene-for-trait 1 × trait 2", or "$G \times T$", interaction.

Scripts to perform the above analyses using Mx (Neale, 1997) can be found at `http://statgen.iop.kcl.ac.uk/gxe/`.

Chapter 5

Epistasis in quantitative trait locus linkage analysis

This Chapter explores a two-locus variance components model of QTL linkage for sibling pairs which incorporates epistasis. For a range of epistatic models the expected variance components and noncentrality parameter per sib pair can be calculated, to indicate the power to detect epistasis. In QTL linkage analysis, additive and epistatic effects are in fact partially confounded: as a result, variance components under submodels are distorted, with two main implications. First, the analysis of a single locus can in fact detect a QTL with no main effect that interacts epistatically with another (unmeasured) locus. That is, single-locus approaches are not necessarily precluding the detection of purely epistatically-interacting loci. Second, because the non-epistatic variance component estimates in submodels can actually reflect epistatic variance, power to formally detect epistasis is low.

5.1 Introduction

Despite growing evidence for the importance of epistasis in the aetiology of complex traits, it has received relatively little methodological treatment in the human quanti-

tative trait loci (QTL) mapping literature. Consequentially, most current approaches are based on single-locus analyses that apparently focus on additive main effects but should miss potentially important epistatic effects. Both molecular genetic studies in nonhuman organisms and quantitative genetic family studies provide compelling motivation to seriously consider epistasis as central in the aetiology of many complex traits and diseases. In a review of QTL research in *Drosophila melanogaster*, Mackay (2001) outlines the genetic landscape that we should expect to encounter in humans: epistatic effects have been found for many major quantitative traits including bristle number (Gurganus et al., 1999), longevity (Leips and Mackay, 2000) and wing shape (Weber et al., 1999). These epistatic effects can be as large as the loci's main effects and can be sex- and environment-specific. Similar conclusions have been drawn from QTL studies in plants (e.g. Jansen, 1996) and mice (e.g. van Wezel et al., 1996). In humans, whether or not epistasis is likely to play a significant role can be estimated by examining familial recurrence rates or covariances (e.g. Risch, 1990). Applying various multilocus models to phenotype data in pedigrees, the largest plausible single-locus contribution and number of other loci implicated can be estimated. For example, such analyses have implicated a significant role of epistasis in schizophrenia and autism (Risch et al., 1999).

If epistasis is indeed so important, efforts to locate QTL using linkage and association strategies should possibly be more commonly framed within a multilocus, epistatic context. However, it is well known that power to detect epistasis in a QTL linkage framework is low. Eaves (1994) found that epistasis considerably reduces the total amount of information available, considering the classical models of duplicate and complementary gene action (described below). In particular, duplicate gene action greatly reduces the total information but is more likely to be detected against an additive genetic background; complementary gene action is, in contrast, virtually indistinguishable from additive effects. Using a variance components frame-

work, Mitchell et al. (1997) report an application of two-locus linkage using extended pedigrees. Two simulated trait loci account for 22% and 0.5% of the trait variance respectively; an epistatic interaction for 14%. Although power to detect the larger single locus was acceptable (70%), power to detect the epistatic effect was only 17%. In contrast, epistatic effects were detected around 6 to 9% of the time between the first locus and an unlinked control marker, suggesting that the test was somewhat liberal in any case, given a nominal false-positive rate of 5%. This Chapter presents analytic results for the power of two-locus sib-pair QTL linkage analysis.

However, rather than simply demonstrating low power to detect epistasis, this Chapter aims to examine the adequacy of single-locus models, i.e. the impact of unmodelled epistasis. The inadequacy of single-locus models is often assumed to be directly proportional to the magnitude of the epistatic variance components under the true model. For example, exploring different epistatic models, researchers asked whether the variance attributable to the marginal, additive effect of loci would be large enough to be detected: Li and Reich (2000) present marginal values for 50 two-locus models, assuming that these provide direct insight into how single-locus linkage would perform when epistasis is actually present. More generally, Frankel and Schork (1996) argue that as epistatic effects are by definition attributable to two or more loci, they will only ever emerge when studying two or more loci jointly. Conversely, considering loci in isolation should be equivalent to looking at their effects averaged over all the other loci with which they may or may not interact. Frankel and Schork (1996) highlight a 'worst-case' scenario in which two genes have no main effects but epistatically influence the trait in a very strong manner (Table 5.1). As the marginals are all equal, the assumption is that such loci would never be detected by single-locus approaches. Whilst, undoubtably, simple single-locus tests of association would fail to detect either locus in this situation, the assumption that tests of linkage will also fail does not necessarily hold, as will be illustrated below.

Geno.	Geno. Freq.	A_1A_1 0.25	A_1A_2 0.50	A_2A_2 0.25	Marginal
B_1B_1	0.25	0	0	1	0.25
B_1B_2	0.50	0	0.50	0	0.25
B_2B_2	0.25	1	0	0	0.25
	Marginal	0.25	0.25	0.25	1.00

Table 5.1: "Two deleterious alleles & two normal alleles bilocus interaction". After Frankel & Schork (1996). This is M_{12} in the models considered below.

It has long been known that linkage strategies are blunt tools for dissecting anything other than the most simple of QTL architecture. In particular, the effects of single-locus and epistatic effects are not necessarily independent from each other. As Mather (1974) stated:

> "In practice the chief consequence of interaction is likely to be to alter the apparent values of D_R and H_R as estimated from variances and covariances."

where D_R and H_R are the polygenic additive and dominance components of variance. This fact has consequences both for the power to detect epistasis and the adequacy of non-epistatic models. Tiwari and Elston (1997b) illustrate how the Haseman-Elston linkage method (Haseman and Elston, 1972) can be adapted to incorporate multiple loci and epistasis. Although the authors note the confounding of epistatic and non-epistatic effects,

> "Therefore, these components of epistatic variance may be better detected in a two-locus model than a one-locus model, *even though they are in principle detectable in the latter.*" [my italics]

they do not evaluate the actual behaviour of the two-locus Haseman-Elston model. Rather, the assumption is made that the power to detect epistasis, relative to the power to detect single locus effects, will be proportional to the relative magnitude of main versus epistatic components of variance.

Effect	Description	Example
4 Main effects	Additive	A_1
6 Two-way interactions	2 Dominance	$A_1 \times A_2$
	4 Additive-additive epistasis	$A_1 \times B_1$
4 Three-way interactions	Additive-dominance epistasis	$A_1 \times B_1 \times B_2$
1 Four-way interaction	Dominance-dominance epistasis	$A_1 \times A_2 \times B_1 \times B_2$

Table 5.2: Partitioning of epistatic interaction effects.

However, for binary disease traits, Culverhouse et al. (2002) recently investigated the class of epistatic models showing no single-locus variation and found that single-locus linkage strategies might succeed where single-locus association would fail, due to the relationship between allele sharing and epistatic genetic effects. This Chapter presents a parallel investigation in the context of quantitative trait variance components linkage analysis.

5.2 Biometrical model of epistasis

The effects of the four alleles present at two loci can be partitioned into main effects and various interaction terms (Cockerham, 1954). The main effect of each allele represents its additive contribution averaged over all alleles with which it may or may not interact. Two-way interactions between alleles at the same locus represent dominance effects. Two-way interactions between two alleles at different loci constitute *additive* $\times$ *additive* ($A \times A$) epistatic effects. The two other higher-order epistatic effects are three-way interactions (i.e. two alleles at one locus and one allele at another locus) creating *additive* $\times$ *dominance* ($A \times D$) or, equally, *dominance* $\times$ *additive* ($D \times A$) epistatic effects and a final interaction term involving all four alleles, *dominance* $\times$ *dominance* ($D \times D$) epistasis. These orders of epistasis are tabulated in Table 5.2 along with some examples in terms of the alleles, '1' and '2', at two diallelic loci, A and B.

All scenarios considered below feature only two loci, with alleles labelled "1" and

	Locus 1		
Locus 2	11	12	22
11	$m + a_1 + a_2 + aa$	$m + d_1 + a_2 + da$	$m - a_1 + a_2 - aa$
12	$m + a_1 + d_2 + ad$	$m + d_1 + d_2 + dd$	$m - a_1 + d_2 - ad$
22	$m + a_1 - a_2 - aa$	$m + d_1 - a_2 - da$	$m - a_1 - a_2 + aa$

Table 5.3: General two-locus epistasis: components of means.

"2" (for variance components, the labels "1" and "2" refer to the locus, rather than the allele, however). Table 5.3 represents this general model with each genotypic mean expressed in terms of main effects and epistatic effects, or what might be called *components of mean.* For example, m is the 'mean' effect; a_1 is the additive genetic value for the first locus; a_2 is the dominance deviation for the second locus. The four epistatic effects are aa, ad, da and dd.

5.2.1 Genetic effects and variance components: a happloid example

Before outlining the calculation of two-locus components of variance, this section considers the same process in a simpler context: a happloid organism. Haploidy is only having a single copy of each chromosome, so there are no dominance effects. The following example is designed to illustrate the relationship between genetic effects and components of variance. For two happloid diallelic loci labelled A and B there are 4 possible two-locus configurations: A_1B_1, A_2B_1, A_1B_2 and A_2B_2. The four genotypic means can be decomposed into the components of genetic effects. The mean effect is m; the additive genetic value for the A locus is a_A; the additive genetic value for the B locus is a_B; the $A \times A$ epistatic effect between loci is aa.

	A_1	A_2
B_1	$m + a_A + a_B + aa$	$m - a_A + a_B - aa$
B_2	$m + a_A - a_B - aa$	$m - a_A - a_B + aa$

In the simplest case, there is only a single main effect at locus A, say $a_A = 1$:

	A_1	A_2
B_1	1	-1
B_2	1	-1

Imagine a population where 10% of individuals have the A_1 allele and 90% have the A_2 allele. Calculating the components of variance involves mean-centering the data, as variances involve deviations from the mean. In this case the population mean would be $1 \times 0.1 + (-1) \times 0.9 = -0.8$. Calculating the variance attributable to locus A averages over any other effects, such as locus B, interaction effects, or residual effects. These averages are the marginal means:

	A_1	A_2	
B_1	1.8	-0.2	0
B_2	1.8	-0.2	0
	1.8	-0.2	0

The variance for frequency-weighted scores is $s^2 = \sum_{i=1}^{n} f_i(x_i - \mu)^2$ where f_i is the relative frequency of the i^{th} of n groups of score x_i and μ is the population mean. Therefore, the variance attributable to the additive effects of the A locus is $0.1 \times 1.8^2 + 0.9 \times (-0.2)^2 = 0.36$. No variance is attributable to locus B [1].

Now introduce an epistatic additive × additive interaction: $m = a_A = a_B = 0$ but $\delta_{AB} = 1$. If alleles B_1 and B_2 are equifrequent, the table of means is:

	A_1	A_2	
B_1	1	-1	-0.8
B_2	-1	1	0.8
	0	0	0

[1] If p and q are allele frequencies and a is the additive genetic value, then the additive genetic variance in diploid organisms is $2pqa^2$. The difference between the two homozygote means is twice the additive genetic value – that is, a_A in this example would be twice the value of a. As diploids have two alleles at each locus, we would expect the additive genetic variance to be twice as large. That is, $2 \times 0.1 \times 0.9 \times 2^2 = 0.72$, which is twice 0.36.

The total mean is already zero, so mean-centering is unnecessary. The B_1 and B_2 marginal genotypic values are nonzero however, due to the unequal allele frequencies of the A_1 and A_2 alleles. That is, the marginal mean of the B_1 allele is $0.1 \times 1 + 0.9 \times (-1) = -0.8$. Because the B_1 and B_2 alleles occur with equal frequency, the effects of $A \times A$ epistasis cancel each other out completely in the marginal means for the A_1 and A_2 alleles.

This has implications on the estimation of the additive variance. The additive single locus variance components are calculated in the same manner as before. As the marginal values of A_1 and A_2 are zero, however, the variance attributable to the main effects of alleles at the A locus is also zero. However, for the B locus, the variance is $0.5 \times (-0.8)^2 + 0.5 \times 0.8^2 = 0.64$. This result may seem counter-intuitive for two reasons: 1) there are no main additive effects at the A or B locus, yet the variance component at locus B is nonzero; 2) the nonzero additive variance for locus B is due to the unequal pattern of allele frequencies at the *other* locus, A.

Nonetheless, this result is quite clear when one considers what would actually be happening in the hypothetical population. The majority of individuals will have the A_2 allele. Therefore, the effect of the epistatic interaction, in the majority of the population, is to decrease B_1 individuals' scores and increase B_2 individuals' scores. The complementary effect, observed in A_1 individuals only occurs 10% of the time. So, on average, having a B_1 allele is associated with having a lower score: that is, there is a main additive effect.

How is the variance component for the epistatic interaction term calculated? As defined, interactions represent deviations from what we would expect given the main effects. Under a strictly additive model, each of the four genotypic means would be the sum of the two corresponding marginal effects, whatever the allele frequencies. The following table represents the four expected genotypic means under a non-epistatic model based on the observed marginal means in the current example:

	A_1	A_2	
B_1	-0.8	-0.8	-0.8
B_2	0.8	0.8	0.8
	0	0	0

This is clearly not what we observe. Returning to the hypothetical population, we now have four categories of individual, representing the four pairwise combinations for alleles: A_1B_1, A_1B_2, A_2B_1 and A_2B_2. Assuming the two loci are in linkage equilibrium, these have frequencies that are the products of the allele frequencies: 0.05, 0.05, 0.45 and 0.45. For the 5% of individuals who have the A_1B_1 combination: they score $1-(-0.8)-0=1.8$ units more than they would have if the two loci operated additively. For the 5% of individuals with the A_1B_2 combination: they score $-1-(-0.8)-0=-0.2$ units more than under an additive model. For the 45% of the population with the A_2B_1 combination: they score $-1-0.8-0=-1.8$ units more. Finally, for the remaining 45% with the A_2B_2 combination: they score $1-0.8-0=0.2$ units more. These values, therefore, represent the epistatic effects in this population: 1.8, -0.2, -1.8 and 0.2. The epistatic variance is simply the variance of these scores: we know the frequencies at which they occur and they are already mean-centered, so the variance can be calculated simply as $0.05\times1.8^2+0.05\times(-0.2)^2+0.45\times(-1.8)^2+0.45\times0.2^2$ which happens to equal 0.36.

So, the above pattern of genotypic means and allele frequencies is associated with additive variances of 0 and 0.64 for loci A and B respectively and an $A\times A$ epistatic variance of 0.36. Although, after a little consideration, the reason for the nonzero additive variance component for locus B is clear, it is not always so easy to intuitively make the connection between main effects and variance components. For example, if we actually add a main effect for locus B so that both a_B and aa equal 1, then the additive variance for locus B actually drops to 0.04. Naturally, the relationship between effects and variance components is often even less transparent in the diploid case, with

eight genetic effects including dominance and four orders of epistatic interaction.

In general, different populations may share identical underlying patterns of gene action but might exhibit quite distinct patterns in terms of the estimated components of variance because of differences in allele frequencies. Furthermore, there is no direct relationship between genetic effects and variance components except in special cases. For example, additive genetic variance will typically be influenced by dominance genetic effects as well as additive genetic effects. The special case for a diallelic locus, is when allele frequencies are equal. In a similar way, epistatic effects also influence additive components of variance. One side effect of this is that it is quite possible to have large epistatic effects but small epistatic variance components. As we shall see, as well as leading to low power to detect such effects, it should also be clear that the parameters estimated in components of variance models are not necessarily good guides to the underlying genetic architecture.

Given this, what utility is there in focusing on components of variance? Historically, classical quantitative genetic models were developed in order to understand polygenic inheritance, during a time when an individual's genotype was unobservable. Consequently, methods related combined effects of alleles and their interactions, rather than individual gene action, to the observable phenotypes. Variance components are natural statistics to describe aggregate statistical effects in specific populations. But even with modern genotyping technology, components of variance are a meaningful metric, relevant to a common goal of research: to answer questions along the lines of "how *important* is this effect?". Not just *how large*, or *how common*, but how important, which will typically involve both the magnitude and frequency of an effect. At the population level, a common moderate effect may well have far greater impact than a rare severe effect. Variance components methods, whilst not *a priori* able to distinguish between these two opposing scenarios, are theoretically able to detect either equally.

5.2.2 Calculation of epistatic variance components

Several derivations for the calculation of variance components associated with epistatic models are available in the literature (e.g. Tiwari and Elston, 1997a). A simple method for the two-locus case is presented in this section.

Let $\mu_{ij|kl}$ be the two-locus genotypic mean for individuals with alleles ij at locus 1 and alleles kl at locus 2. Indices i, j, k and l take the values "1" or "2" representing the alternate alleles at the diallelic loci; genotypes are unordered. The allele frequencies of the "1" allele are p_1 and p_2 for the first and second locus respectively (allele "2" has frequency $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$). The recombination fraction between the two loci is θ, typically set at 0.5, to indicate the loci are unlinked loci. Residual variance shared between siblings is σ_S^2, nonshared residual variance is σ_N^2. If the total variance is assumed to be 1, then the total QTL variance is $1 - \sigma_S^2 - \sigma_N^2$.

The following calculations follow the logic of the haploid example, if not the detail. The grand mean is calculated as the weighted sum of all genotypic means,

$$\mu_{..|..} = p_1^2 p_2^2 \mu_{11|11} + p_1^2 2p_2 q_2 \mu_{11|12} + p_1^2 q_2^2 \mu_{11|22} + \cdots + q_1^2 q_2^2 \mu_{22|22}$$

and all genotypic means are re-expressed as deviations from this grand mean.

The marginal genotypic effects are calculated as follows. For the marginal genotypic effect of the "11" genotype at locus 1, the sums of the consistent joint genotypic means are weighted by the probability of them occurring conditional on the genotype being "11" at locus 1. That is,

$$\mu_{11|..} = p_2^2 \mu_{11|11} + 2p_2 q_2 \mu_{11|12} + q_2^2 \mu_{11|22}$$

and a similar logic applies to all the other marginal genotypic means, as well as the

marginal allelic means. For example,

$$\begin{aligned}\mu_{1.|..} &= p_1p_2^2\mu_{11|11} + 2p_1p_2q_2\mu_{11|12} + p_1q_2^2\mu_{11|22} \\ &\quad + q_1p_2^2\mu_{12|11} + 2q_1p_2q_2\mu_{12|12} + q_1q_2^2\mu_{12|22}\end{aligned}$$

Considering next the epistatic effects, two of the four additive × additive interactive effects, for each pair of alleles between loci, are illustrated below:

$$\mu_{1.|1.} = \mu_{11|11}p_1p_2 + \mu_{11|12}p_1q_2 + \mu_{12|11}q_1p_2 + \mu_{12|12}q_1q_2$$

$$\mu_{1.|2.} = \mu_{11|12}p_1p_2 + \mu_{11|22}p_1q_2 + \mu_{12|12}q_1p_2 + \mu_{12|22}q_1q_2$$

and for additive × dominance (and dominance × additive) interactions, four of the twelve terms are

$$\mu_{1.|11} = p_1\mu_{11|11} + q_1\mu_{12|11}$$

$$\mu_{2.|11} = p_1\mu_{12|11} + q_1\mu_{22|11}$$

$$\mu_{11|1.} = p_2\mu_{11|11} + q_2\mu_{11|12}$$

$$\mu_{11|2.} = p_2\mu_{11|12} + q_2\mu_{11|22}$$

Having calculated these marginal means, we are in a position to calculate the "deviation" scores for the different types of interaction, from which the variance components are calculated. That is, the following quantities represent the effects of having a particular combination of alleles over and above what one would expect without allowing for that particular type of interaction, in that population (as they are allele frequency dependent, as we saw in the haploid example). They follow a hierarchical pattern, in that three-way interactions are considered in the presence of all possible two-way interactions, and so on. We will call these quantities population interaction deviations. For the main effects of alleles, these quantities are labelled α and are equal to the

corresponding marginal allelic mean.[2]

Representing the effects of dominance, (i.e. two-way interactions between alleles at the same locus), the population interaction deviations are $\gamma_{ij|\diamond\diamond}$ for the ij genotype at the first locus; and $\gamma_{\diamond\diamond|kl}$ for the kl genotype at the second locus.

$$\begin{aligned}
\gamma_{11|\diamond\diamond} &= \mu_{11|..} - 2\alpha_{1|\diamond} \\
\gamma_{12|\diamond\diamond} &= \mu_{12|..} - \alpha_{1|\diamond} - \alpha_{2|\diamond} \\
\gamma_{22|\diamond\diamond} &= \mu_{22|..} - 2\alpha_{2|\diamond} \\
\gamma_{\diamond\diamond|11} &= \mu_{..|11} - 2\alpha_{\diamond|1} \\
\gamma_{\diamond\diamond|12} &= \mu_{..|12} - \alpha_{\diamond|1} - \alpha_{\diamond|2} \\
\gamma_{\diamond\diamond|22} &= \mu_{..|22} - 2\alpha_{\diamond|2}
\end{aligned}$$

For $A \times A$ interaction, $\epsilon_{i\diamond|j\diamond}$ is the deviation for the i^{th} allele at locus A and the j^{th} allele at locus B. For example,

$$\epsilon_{1\diamond|1\diamond} = \mu_{1.|1.} - \alpha_{1|\diamond} - \alpha_{\diamond|1}$$

$$\epsilon_{1\diamond|2\diamond} = \mu_{1.|2.} - \alpha_{1|\diamond} - \alpha_{\diamond|2}$$

For higher-order epistatic interactions, $\epsilon_{ij|k\diamond}$ is the dominance $\times$ additive deviation for the ij genotype at locus A and the k allele at locus B. For example,

$$\begin{aligned}
\epsilon_{11|1\diamond} &= \mu_{11|1.} - 2\mu_{1.|1.} - \gamma_{11|\diamond\diamond} - 2\alpha_{1|\diamond} - \alpha_{\diamond|1} \\
\epsilon_{11|2\diamond} &= \mu_{11|2.} - 2\mu_{1.|2.} - \gamma_{11|\diamond\diamond} - 2\alpha_{1|\diamond} - \alpha_{\diamond|1} \\
\epsilon_{12|1\diamond} &= \mu_{12|1.} - \mu_{1.|1.} - \mu_{2.|1.} - \gamma_{12|\diamond\diamond} - \alpha_{1|\diamond} - \alpha_{2|\diamond} - \alpha_{\diamond|1}
\end{aligned}$$

[2]The notation has changed, so that the dot (.) symbol is replaced by the $\diamond$ symbol in the notation, to represent the fact that these are not marginal means (i.e. typically the dot indicates a weighted summation). The diamonds are added in order to make the positional notation clear.

Likewise, $\epsilon_{i\circ|kl}$ is the additive × dominance deviation for the kl genotype at locus B and the i allele at locus A. Finally, dominance × dominance interactions are represented by $\epsilon_{ij|kl}$ for the ij genotype at locus A and the kl genotype at locus B; the first two of the nine terms are:

$$\begin{aligned}
\epsilon_{11|11} &= \mu_{11|11} - 2\epsilon_{11|1\circ} - 2\epsilon_{1\circ|11} - 4\epsilon_{1\circ|1\circ} - \gamma_{11|\circ\circ} - \gamma_{\circ\circ|21} - 2\alpha_{1|\circ} - 2\alpha_{\circ|1} \\
\epsilon_{11|12} &= \mu_{11|12} - \epsilon_{11|1\circ} - \epsilon_{11|2\circ} - 2\epsilon_{1\circ|12} \\
&\quad -2\epsilon_{1\circ|1\circ} - 2\epsilon_{1\circ|2\circ} - \gamma_{11|\circ\circ} - \gamma_{\circ\circ|12} - 2\alpha_{1|\circ} - \alpha_{\circ|1} - \alpha_{\circ|2}
\end{aligned}$$

and the dominance × dominance term for the double-heterozygote is given by

$$\begin{aligned}
\epsilon_{12|12} &= \mu_{12|12} - \epsilon_{12|1\circ} - \epsilon_{12|2\circ} - \epsilon_{1\circ|12} - \epsilon_{2\circ|12} - \epsilon_{1\circ|1\circ} - \epsilon_{1\circ|2\circ} - \epsilon_{2\circ|1\circ} - \epsilon_{2\circ|2\circ} \\
&\quad -\gamma_{12|\circ\circ} - \gamma_{\circ\circ|12} - \alpha_{1|\circ} - \alpha_{2|\circ} - \alpha_{\circ|1} - \alpha_{\circ|2}
\end{aligned}$$

The associated variance components are then calculated as follows

$$\begin{aligned}
\sigma^2_{A1} &= 2(p_1(\alpha_{1|\circ})^2 + q_1(\alpha_{2|\circ})^2) \\
\sigma^2_{A2} &= 2(p_2(\alpha_{\circ|1})^2 + q_2(\alpha_{\circ|2})^2) \\
\sigma^2_{D1} &= p_1^2(\gamma_{11|\circ\circ})^2 + 2p_1q_1(\gamma_{12|\circ\circ})^2 + q_1^2(\gamma_{22|\circ\circ})^2 \\
\sigma^2_{D2} &= p_2^2(\gamma_{\circ\circ|11})^2 + 2p_2q_2(\gamma_{\circ\circ|12})^2 + q_2^2(\gamma_{\circ\circ|22})^2
\end{aligned}$$

whilst the epistatic components of variance are

$$\begin{aligned}
\sigma^2_{AA} &= 4\Big(p_1p_2(\epsilon_{1\circ|1\circ})^2 + p_1q_2(\epsilon_{1\circ|2\circ})^2 + q_1p_2(\epsilon_{2\circ|1\circ})^2 + q_1q_2(\epsilon_{2\circ|2\circ})^2\Big) \\
\sigma^2_{AD} &= 2\Big(p_1p_2^2(\epsilon_{11|1\circ})^2 + q_1p_2^2(\epsilon_{11|2\circ})^2 + 2p_1p_2q_2(\epsilon_{12|1\circ})^2 \\
&\quad +2q_1p_2q_2(\epsilon_{12|2\circ})^2 + p_1q_2^2(\epsilon_{22|1\circ})^2 + q_1q_2^2(\epsilon_{22|2\circ})^2\Big)
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{DA} &= 2\Big(p_2p_1^2(\epsilon_{1\circ|11})^2 + q_2p_1^2(\epsilon_{2\circ|11})^2 + 2p_2p_1q_1(\epsilon_{1\circ|12})^2 \\
&\quad +2q_2p_1q_1(\epsilon_{2\circ|12})^2 + p_2q_1^2(\epsilon_{1\circ|22})^2 + q_2q_1^2(\epsilon_{2\circ|22})^2\Big) \\
\sigma^2_{DD} &= p_1^2p_2^2(\epsilon_{11|11})^2 + 2p_1^2p_2q_2(\epsilon_{11|12})^2 + p_1^2q_2^2(\epsilon_{11|22})^2 \\
&\quad +2p_1q_1p_2^2(\epsilon_{12|11})^2 + 4p_1q_1p_2q_2(\epsilon_{12|12})^2 + 2p_1q_1q_2^2(\epsilon_{12|22})^2 \\
&\quad +q_1^2p_2^2(\epsilon_{22|11})^2 + 2q_1^2p_2q_2(\epsilon_{22|12})^2 + q_1^2q_2^2(\epsilon_{22|22})^2
\end{aligned}$$

Finally, these variance components are standardised; first the total QTL variance is calculated as the sum of all eight QTL variance components

$$\sigma^2_T = \sigma^2_{A1} + \sigma^2_{A2} + \sigma^2_{D1} + \sigma^2_{D2} + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DA} + \sigma^2_{DD}$$

and then the QTL variance components are recalculated such that they will sum to $1 - \sigma^2_S - \sigma^2_N$. For example, the additive variance for the first locus is recalculated as

$$\sigma^2_{A1} = \frac{\sigma^2_{A1}(1 - \sigma^2_S - \sigma^2_N)}{\sigma^2_T}$$

The components of variance, calculated for a specific epistatic model and set of allele frequencies can then be used in determining the expected noncentrality parameter of the linkage test, which is solely a function of the variance components.

Tiwari and Elston (1997a) derived the expected components of variance for quantitative two-locus models, in order to facilitate the exploration of power issues. Using a method involving partial derivatives of the population mean (Kojima, 1959), formulae for calculating the components of variance from a matrix of genotypic means and allele frequencies for two unlinked, diallelic loci. The procedure presented above provides identical results. Calculating the magnitude of the variance components under the true model is only the first step in the assessment of epistasis in QTL linkage, however: Tiwari and Elston (1997a) and in following papers (Tiwari and Elston,

1997b, 1998) essentially do not go beyond this point.

5.2.3 Specific models of epistasis

In order to illustrate the behaviour of the two-locus QTL linkage model, we focus on a specific set of models with a limited range of phenotypic means, typically 0 and 1. Li and Reich (2000) enumerated and attempted to classify all possible two-locus models assuming a binary trait. Although there are $2^9 = 512$ different two-locus models (i.e. 9 genotypic means, each with 2 possible states) only 50 of the 512 models are unique. For example, two models are equivalent if one can be obtained by switching affection status for all 9 cells. The current work will focus on 13 two-locus models, M_1 to M_{13} as below. Many of these models have been previously examined in the context of binary traits (Neuman and Rice, 1992; Schork, 1993). Each matrix of genotypic means corresponds to the 9 cells in Table 5.3.

$$M_1 = \begin{pmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \end{pmatrix} M_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} M_3 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} M_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} M_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

$$M_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix} M_8 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} M_9 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

$$M_{10} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} M_{11} = \begin{pmatrix} 1 & \mathrm{x} & 0 \\ \mathrm{x} & 0 & \mathrm{x} \\ 0 & \mathrm{x} & 1 \end{pmatrix} M_{12} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0.5 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

$$M_{13} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Models M_1 to M_3 are single-locus models, included for comparison. M_1 is an additive single-locus model. Model M_2 represents a recessive trait but also a dominant trait if 'affection status' were reversed. If the frequency of the A_1 allele is high, then it represents a rare recessive disorder, where "1" represents caseness, for example. M_3 could be referred to as a single-locus "interference" model, or "overdominance".

Model M_4 represents complementary gene action for a recessive trait. Alternatively, with affection status swapped, M_4 would represent duplicate gene action for a dominant trait. All models will be investigated under a range of allele frequencies, so both rare recessive (frequency of "2" alleles low and 1 is "affected") and rare dominant (frequency of "1" alleles low and 0 is "affected") will be covered. From now on, these symmetries will not be noted for the basic models. M_5 represents the case of dominant × recessive complementary gene action for loci A and B respectively. M_6, a "threshold" model requires that at least 3 of the "2" alleles are present; which loci they originate from is irrelevant. M_7 is a "modifying-effects" model in that only a slight modification of a single-locus recessive model (locus B) has resulted in epistasis. M_8 is the dominant × dominant complementary gene action model. M_9 is sometimes called the "XOR" (exclusive OR) model (and has been implicated in the genetics of handedness).

Models M_{10} through M_{13} represent some of the more extreme possible cases of epistasis. M_{10} is sometimes referred to as the "checkerboard" pattern. M_{11} represents a "balance–imbalance" model of epistasis. The x parameter could either be 0, 0.5 or 1, representing "recessive", "additive" and "dominant" forms of this interaction respectively. A similar model, M_{12}, represents the example given by Frankel and

Schork (1996) when both loci have equal allele frequency, as we shall review in more depth below. In M_{13} only double heterozygotes are affected, which corresponds to only the $D \times D$ genetic effect being nonzero.

The associated QTL variance components and noncentrality parameter (see below) are first calculated under the full model including all epistatic terms. Second, the *apparent* variance components under various nested submodels are calculated, along with the associated drop in the noncentrality parameter. The apparent variance components indicate what we would expect to find if we assumed a certain model (i.e. no epistasis) which is different from reality (i.e. epistasis). This will enable exploration of model misspecification effects. As the test for epistasis is based on the difference in fit between (1) a model with single locus effects and epistatic effects and (2) a model with only single locus effects, this procedure also enables us to investigate the power of the variance components method to detect epistasis.

5.3 Epistasis and quantitative trait loci

Tiwari and Elston (1997b) illustrate how the Haseman-Elston linkage method can be adapted to incorporate multiple loci and epistasis. For a single locus, the regression coefficient estimates $-2(1-2\theta)^2\sigma_A^2$ where θ is the recombination fraction between marker and trait locus and σ_A^2 is the additive QTL variance. Since the dependent variable is the squared mean-corrected sibling trait difference, a significantly negative regression slope is taken to be evidence for linkage. No model for the mode of inheritance need be specified prior to performing the analysis.

The two-locus extension considers two unlinked trait loci and two marker loci (each one linked to one of the trait loci but also unlinked to each other). At the markers, assuming linkage equilibrium between marker and trait locus, as well as π measured for both loci, f_1 is the probability of sharing precisely 1 allele IBD at the

first marker and f_2 is the probability of sharing 1 allele IBD at the second locus. The Haseman-Elston regression is based on

$$\begin{aligned} E(X|\pi_1, f_1, \pi_2, f_2) &= \alpha + \beta_1\pi_1 + \beta_2\pi_2 + \delta_1 f_1 + \delta_2 f_2 \\ &+ \gamma_{AA}\pi_1\pi_2 + \gamma_{AD}\pi_1 f_2 + \gamma_{DA} f_1\pi_2 + \gamma_{DD} f_1 f_2 \end{aligned}$$

The β_1 and δ_1 regression coefficients estimate the following quantities (where $\Psi_i = \theta_i^2 + (1-\theta_i)^2$ and θ_i is the QTL-marker recombination fraction for marker-locus pair i):

$$\begin{aligned} \beta_1 &= 2(1-2\Psi_1)[\sigma^2_{A1} + \sigma^2_{D1} + (1-\Psi_2)(\sigma^2_{AA} + \sigma^2_{DA}) + (1-\Psi_2)^2(\sigma^2_{AD} + \sigma^2_{DD})] \\ \delta_1 &= (1-2\Psi_1)^2[\sigma^2_{D1} + (1-\Psi_2)\sigma^2_{DA} + (1-\Psi_2)^2\sigma^2_{DD}] \end{aligned}$$

If the second marker is unlinked ($\Psi_2 = 0.5$) and there are no epistatic terms, then β_1 reduces to the original Haseman-Elston coefficient: $2(1-2\Psi_1)\sigma^2_A = -2(1-2\theta)^2\sigma^2_A$. If both markers are completely linked to their respective traits (i.e. $\Psi_1 = \Psi_2 = 1$) then

$$\begin{aligned} \beta_1 &= -2(\sigma^2_{A1} + \sigma^2_{D1}) \\ \beta_2 &= -2(\sigma^2_{A2} + \sigma^2_{D2}) \\ \delta_1 &= \sigma^2_{D1} \\ \delta_2 &= \sigma^2_{D2} \\ \gamma_{AA} &= -2(\sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DA} + \sigma^2_{DD}) \\ \gamma_{AD} &= \sigma^2_{AD} + \sigma^2_{DD} \\ \gamma_{DA} &= \sigma^2_{DA} + \sigma^2_{DD} \\ \gamma_{DD} &= -\frac{1}{2}\sigma^2_{DD} \end{aligned}$$

The authors note the confounding of epistatic and non-epistatic effects: to quote

Tiwari and Elston (1997b):

> "In general, when $\Psi_2 = 0.5$, β_1 and δ_1 include additional epistatic variance terms, so that values of β_1 and δ_1 will be inflated if these variances are not zero. However, note that the coefficient of σ^2_{AA} and σ^2_{DA} are $(1 - 2\Psi_1)$ in the expression for β_1 and that of σ^2_{AD} and σ^2_{DD} are $\frac{1}{2}(1 - 2\Psi_1)$, which are smaller in magnitude than their respective coefficients in γ_{AA} if $\Psi_2 = 1$. Therefore, these components of epistatic variance may be better detected in a two-locus model than a one-locus model, even though they are in principle detectable in the latter."

In this case, if epistatic components of variance exist, we would expect γ_{AA} and γ_{DD} to be significantly below zero, and γ_{AD} and γ_{DA} to be significantly above zero. The significance of the γ coefficients would then provide specific tests for various types of epistasis. Chapter 8 presents an alternative formulation of the Haseman-Elston regression model incorporating two-locus epistasis. In particular, the new approach is more powerful and allows specific tests of individual epistatic variance components.

5.3.1 QTL variance components linkage model

Variance components models allow tests for linkage in sibships of any size, incorporating additive and dominance effects at QTL. The covariance structure is modelled in terms of the proportion of alleles shared identical by descent (IBD) denoted π and the probability of complete IBD sharing, z, at the candidate locus for every sib pair in the sibship. Extending the basic model to two interacting loci, trait variance is decomposed into eight QTL components and two residual components. Assuming random mating, and that the two loci are in linkage equilibrium, the trait variance is

$$Var(X) = \sigma^2_{A1} + \sigma^2_{A2} + \sigma^2_{D1} + \sigma^2_{D2} + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DA} + \sigma^2_{DD} + \sigma^2_S + \sigma^2_N$$

For convenience, the trait variance is fixed to 1. The expected sibling correlation r is a function of components of variance shared between siblings. Under the alternate hypothesis of linkage, this sharing will depend on IBD status at both loci, such that the expected sibling correlation is

$$\begin{aligned} r_L &= \hat{\pi}_1\sigma^2_{A1} + \hat{\pi}_2\sigma^2_{A2} + \hat{z}_1\sigma^2_{D1} + \hat{z}_2\sigma^2_{D2} + \hat{\pi}_1\hat{\pi}_2\sigma^2_{AA} \\ &\quad + \hat{\pi}_1\hat{z}_2\sigma^2_{AD} + \hat{z}_1\hat{\pi}_2\sigma^2_{DA} + \hat{z}_1\hat{z}_2\sigma^2_{DD} + \sigma^2_S \end{aligned}$$

where the two-locus allele-sharing variables are simply the products of the corresponding single-locus variables. Under the null, the IBD variables take their expected values so the expected sibling correlation is

$$\begin{aligned} r_N &= E[\hat{\pi}_1]\sigma^2_{A1} + E[\hat{\pi}_2]\sigma^2_{A2} + E[\hat{z}_1]\sigma^2_{D1} + E[\hat{z}_2]\sigma^2_{D2} \\ &\quad + E[\hat{\pi}_1\hat{\pi}_2]\sigma^2_{AA} + E[\hat{\pi}_1\hat{z}_2]\sigma^2_{AD} + E[\hat{z}_1\hat{\pi}_2]\sigma^2_{DA} + E[\hat{z}_1\hat{z}_2]\sigma^2_{DD} + \sigma^2_S \end{aligned}$$

where the two-locus allele-sharing expectations also depend on the recombination fraction between the two loci. If θ is the recombination fraction and $\Psi = \theta^2 + (1-\theta)^2$ then the joint probabilities of allele-sharing at both loci (0, 1 or 2 alleles, counting right across columns and down across rows) is given in the 3×3 matrix (Haseman and Elston, 1972)

$$\mathbf{P} = \begin{pmatrix} \frac{\Psi^2}{4} & \frac{\Psi(1-\Psi)}{2} & \frac{(1-\Psi)^2}{4} \\ \frac{\Psi(1-\Psi)}{2} & \frac{1-2\Psi(1-\Psi)}{2} & \frac{\Psi(1-\Psi)}{2} \\ \frac{(1-\Psi)^2}{4} & \frac{\Psi(1-\Psi)}{2} & \frac{\Psi^2}{4} \end{pmatrix}$$

which, for unlinked loci ($\theta = 0.5$), reduces to

$$\begin{pmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{pmatrix}$$

Similarly, eight **S** matrices define the allele-sharing variables π_1, z_1, π_2, z_2, $\pi_1\pi_2$, $\pi_1 z_2$, $z_1\pi_2$ and $z_1 z_2$. Labelled $\mathbf{S}_1$ to $\mathbf{S}_8$, each is a 3×3 matrix with elements corresponding to the 9 joint IBD configurations.

$$\mathbf{S}_1 = \begin{pmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \end{pmatrix} \quad \mathbf{S}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\mathbf{S}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0.5 & 0.5 & 0.5 \\ 1 & 1 & 1 \end{pmatrix} \quad \mathbf{S}_4 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

$$\mathbf{S}_5 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.25 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix} \quad \mathbf{S}_6 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0.5 & 1 \end{pmatrix}$$

$$\mathbf{S}_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 1 \end{pmatrix} \quad \mathbf{S}_8 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The expected value of allele-sharing variable i is $\sum_j \sum_k ([\mathbf{P}]_{jk} \times [\mathbf{S}_i]_{jk})$. For two unlinked loci, this gives

$$r_N = \frac{\sigma^2_{A1}}{2} + \frac{\sigma^2_{A2}}{2} + \frac{\sigma^2_{D1}}{4} + \frac{\sigma^2_{D2}}{4} + \frac{\sigma^2_{AA}}{4} + \frac{\sigma^2_{AD}}{8} + \frac{\sigma^2_{DA}}{8} + \frac{\sigma^2_{DD}}{16} + \sigma^2_S.$$

5.3.2 Calculation of the noncentrality parameter (NCP)

Sham et al. (2000b) showed that the power of the QTL linkage test is given by the expectation for the noncentrality parameter conditional on the true model parameters. For unselected samples, the expected noncentrality parameter of the likelihood ratio test is

$$\lambda_L = E(2\ln L_L) - E(2\ln L_N) = -E(\ln|\Sigma_L|) + \ln|\Sigma_N|$$

which represents the expected contribution to the likelihood-ratio test statistic from each sibship, where Σ_L and Σ_N are the expected covariance matrices under the alternate and the null respectively. Ignoring a constant that cancels, $E(\ln|\Sigma_L|) = \sum_{i=1}^{M} p_i \ln|\Sigma_i|$ where p_i is the probability of the i^{th} of M IBD configurations, for which Σ_i is the associated expected covariance matrix. In this present context, all the Σ matrices are 2×2 correlation matrices, so the determinants will have the form $(1 - r^2)$ where r is the expected sibling correlation.

If $\mathbf{R}$ is a 3×3 matrix of correlations conditional on the 9 joint IBD configurations

$$\begin{aligned}\mathbf{R} &= \sigma^2_{A1} \otimes \mathbf{S}_1 + \sigma^2_{D1} \otimes \mathbf{S}_2 + \sigma^2_{A2} \otimes \mathbf{S}_3 + \sigma^2_{D2} \otimes \mathbf{S}_4 + \\ &\quad \sigma^2_{AA} \otimes \mathbf{S}_5 + \sigma^2_{AD} \otimes \mathbf{S}_6 + \sigma^2_{DA} \otimes \mathbf{S}_7 + \sigma^2_{DD} \otimes \mathbf{S}_8\end{aligned}$$

then the expected NCP per sibling pair is

$$\begin{aligned}\lambda &= -E(\ln|\Sigma_L|) + \ln|\Sigma_N| \\ &= -\sum_{i=0}^{2}\sum_{j=0}^{2}\left([\mathbf{P}]_{ij}\ln(1-[\mathbf{R}]^2_{ij})\right) + \ln(1-r_N^2)\end{aligned}$$

and the sample NCP is $N\lambda$ where N is the number of pairs.

An example

As an exàmple, consider epistatic model M_2 where $p_1 = 0.2$ and $p_2 = 0.5$ and the loci are unlinked; all QTL effects account for 10% of the trait variance. After calculating the genetic effects from the matrix of genotypic means and allele frequencies, the following standardised components of variance are obtained:

σ^2_{A1}	0.0004	σ^2_{D1}	0.0009
σ^2_{A2}	0.0632	σ^2_{D2}	0.0316
σ^2_{AA}	0.0009	σ^2_{AD}	0.0004
σ^2_{DA}	0.0018	σ^2_{DD}	0.0009

Given that $\sigma^2_S = 0.2$, the nine expected sibling correlations can be calculated from these variance components. For example, the correlation for individuals sharing 1 allele IBD at locus A and 2 alleles IBD at locus B is

$$\frac{0.0004}{2} + 0.0632 + 0.0316 + \frac{0.0009}{2} + \frac{0.0004}{2} + 0.2 = 0.2956$$

which can be seen as the $[3, 2]$ element [*row,column*] of the full matrix of expected correlations

$$\begin{bmatrix} 0.2000 & 0.2002 & 0.2013 \\ 0.2316 & 0.2320 & 0.2342 \\ 0.2947 & 0.2956 & 0.3000 \end{bmatrix}$$

whilst under the null the correlation is 0.2405:

$$\begin{aligned} r_N \;=\; & \frac{1}{2} \times 0.0004 + \frac{1}{4} \times 0.0009 + \frac{1}{2} \times 0.0632 + \frac{1}{4} \times 0.0316 + \frac{1}{4} \times 0.0009 + \\ & \frac{1}{8} \times 0.0004 + \frac{1}{8} \times 0.0018 + \frac{1}{16} \times 0.0009 + 0.2 = 0.2405 \end{aligned}$$

as the loci are unlinked. Using the IBD configuration probabilities for two unlinked loci, the exact NCP per sibling pair can then be calculated (note: rounding the corre-

lations to four decimal places will make this hand-calculation a slight approximation)

$$\begin{aligned}\lambda &= -\frac{1}{16}\ln(1-0.2000^2) - \frac{1}{8}\ln(1-0.2002^2) - \frac{1}{16}\ln(1-0.2013^2) - \frac{1}{8}\ln(1-0.2316^2) \\ &\quad -\frac{1}{4}\ln(1-0.2320^2) - \frac{1}{8}\ln(1-0.2342^2) - \frac{1}{16}\ln(1-0.2947^2) - \frac{1}{8}\ln(1-0.2956^2) \\ &\quad -\frac{1}{16}\ln(1-0.3000^2) + \ln(1-0.2405^2) \\ &\approx 0.0015\end{aligned}$$

This value represents the NCP per sib pair. For a sample of 2000 unselected sib pairs, therefore, we would expect the full model to have a NCP of $2000 \times 0.0015 \approx 2.94$.

5.3.3 Approximation

Sham et al. (2000b) also show that the expected NCP for a sib pair can also be approximately expressed in terms of the expected variance of the sib correlation which in the present context equals $\mathbf{b'Cb}$ where $\mathbf{b}$ is a vector of the eight QTL variance components and $\mathbf{C}$ is the covariance matrix of the two-locus allele-sharing variables where the $[i^{th}, j^{th}]$ element of $\mathbf{C}$ is

$$[\mathrm{C}]_{ij} = \sum_{k=0}^{2}\sum_{l=0}^{2}[\mathrm{P}]_{kl}[\mathrm{S}_i]_{kl}[\mathrm{S}_j]_{kl} - \left[\left(\sum_{k=0}^{2}\sum_{l=0}^{2}[\mathrm{P}]_{kl}[\mathrm{S}_i]_{kl}\right)\left(\sum_{k=0}^{2}\sum_{l=0}^{2}[\mathrm{P}]_{kl}[\mathrm{S}_j]_{kl}\right)\right]$$

which for unlinked loci gives

$$\mathbf{C} = \begin{pmatrix} \frac{1}{8} & & & & & & & \\ 0 & \frac{1}{8} & & & & & & \\ \frac{1}{8} & 0 & \frac{3}{16} & & & & & \\ 0 & \frac{1}{8} & 0 & \frac{3}{16} & & & & \\ \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{1}{16} & \frac{5}{64} & & & \\ \frac{1}{32} & \frac{1}{16} & \frac{1}{32} & \frac{3}{32} & \frac{1}{16} & \frac{5}{64} & & \\ \frac{1}{16} & \frac{1}{32} & \frac{3}{32} & \frac{1}{32} & \frac{1}{16} & \frac{3}{64} & \frac{5}{64} & \\ \frac{1}{32} & \frac{1}{32} & \frac{3}{64} & \frac{3}{64} & \frac{3}{64} & \frac{7}{128} & \frac{7}{128} & \frac{15}{256} \end{pmatrix}$$

Expanding the matrix notation, after some algebraic rearrangement, gives the approximation to the NCP:

$$\begin{aligned} \lambda_L = & \frac{(\sigma^2_{A1})^2 + (\sigma^2_{A2})^2}{8} + \frac{3((\sigma^2_{D1})^2 + (\sigma^2_{D2})^2)}{16} + \frac{\sigma^2_{A1}\sigma^2_{D1} + \sigma^2_{A2}\sigma^2_{D2}}{4} \\ & + \frac{5((\sigma^2_{AA})^2 + (\sigma^2_{AD})^2 + (\sigma^2_{DA})^2)}{64} + \frac{15(\sigma^2_{DD})^2}{256} \\ & + \frac{\sigma^2_{AA}}{8}\left[\sigma^2_{A1} + \sigma^2_{A2} + \sigma^2_{D1} + \sigma^2_{D2} + \sigma^2_{AD} + \sigma^2_{DA} + \frac{3\sigma^2_{DD}}{4}\right] \\ & + \frac{\sigma^2_{DD}}{16}\left[\sigma^2_{A1} + \sigma^2_{A2} + \frac{3(\sigma^2_{D1} + \sigma^2_{D2})}{2} + \frac{7(\sigma^2_{AD} + \sigma^2_{DA})}{4}\right] \\ & + \frac{\sigma^2_{AD}}{16}\left[\sigma^2_{A1} + 2\sigma^2_{A2} + \sigma^2_{D1} + 3\sigma^2_{D2} + \frac{3\sigma^2_{DA}}{2}\right] \\ & + \frac{\sigma^2_{DA}}{16}\left[2\sigma^2_{A1} + \sigma^2_{A2} + 3\sigma^2_{D1} + \sigma^2_{D2}\right] \end{aligned}$$

which is the approximate NCP per sib pair expressed in terms of the true population variance components. This illustrates that the power of the linkage test depends upon the square of the additive QTL variance. Although epistatic variance components contribute to the NCP, their contributions are more greatly attenuated.

5.3.4 Apparent variance components under nested submodels

So far we have only considered the full model NCP under the true parameter values. In a test of epistasis, the full model is compared to a nested submodel that includes additive effects for both loci but no epistatic components. As mentioned above, given values for the variance components under the full model, it is possible to calculate the *apparent* variance components associated with submodels, and thereby calculate the submodel fit. This procedure will also allow exploration of the adequacy of single-locus approximations, by comparing the single-locus model against the null model.

Nine submodels are considered, as outlined in Table 5.4. A • symbol indicates that the term is estimated in the model, the ○ symbol that it is fixed to zero. Submodels 1 and 2 include epistatic components; submodels 3 and 4 are two-locus models assuming interlocus additivity (no epistasis); submodels 5 to 8 are all single locus models, for each locus separately, both with and without a dominance term; submodel 9 represents no QTL effects for either locus. Various likelihood-ratio tests for specific components of variance can be constructed. For example, comparing the full model against submodel 9 tests for any effect from both QTL; the full model against submodel 3 tests for any epistasis; the full model against submodel 2 tests for only the higher order forms of epistasis.

Two separate methods are employed to calculate submodel apparent variance components: a least squares approximation and a full maximum likelihood method. Given the vector $\mathbf{b}$ of eight variance components under the full model and the 8×8 covariance matrix of IBD sharing variables, $\mathbf{C}$, the vector of covariances between the 8 IBD sharing variables and the mean-centred trait cross-products is $\mathbf{d} = \mathbf{Cb}$. The apparent components of variance under a submodel, e.g. with the $D \times D$ component dropped, are $\mathbf{b_S} = (\mathbf{C_S})^{-1}\mathbf{d_S}$ where $\mathbf{C_S}$ is a 7×7 submatrix of $\mathbf{C}$ and $\mathbf{d_S}$ is a 7 element subvector of $\mathbf{d}$. Although the parameter σ^2_S has zero variance as it is not a fixed effect,

Model	σ^2_{A1}	σ^2_{D1}	σ^2_{A2}	σ^2_{D2}	σ^2_{AA}	σ^2_{AD}	σ^2_{DA}	σ^2_{DD}	σ^2_S	σ^2_N
Full	●	●	●	●	●	●	●	●	●	●
1	●	●	●	●	●	●	●	○	●	●
2	●	●	●	●	●	○	○	○	●	●
3	●	●	●	●	○	○	○	○	●	●
4	●	○	●	○	○	○	○	○	●	●
5	●	●	○	○	○	○	○	○	●	●
6	○	○	●	●	○	○	○	○	●	●
7	●	○	○	○	○	○	○	○	●	●
8	○	○	●	○	○	○	○	○	●	●
9	○	○	○	○	○	○	○	○	●	●

Table 5.4: Variance components estimated under the full and nested submodels.

and so does not feature in the matrix $\mathbf{C}$, the apparent value of σ^2_S can be easily calculated by constraining submodels to give the same expected correlation as the full model. Alternatively, using maximum likelihood (ML) estimation procedures, the various submodels outlined in Table 5.4 can be fitted to the correlational structure expected under the full, true model. The two methods give similar results.

5.4 Results

Table 5.5 shows the full model expected NCP for all models, at varying allele frequencies. In all cases the total QTL variance of both loci combined accounts for the same proportion of the trait variance (10%). The column "NCP" is the expected NCP relative to that expected for model M_1 with equal allele frequencies. The single-locus model M_3 provides the strongest evidence for linkage, almost 1.5 times that of the M_1 baseline. All full model QTL variance loads onto the single-locus dominance components in model M_3. Most epistatic models show a marked reduction in the total amount of linkage information available. The most extreme epistatic models typically result in half the amount of information relative to M_1, even under the full model when the epistasis is modelled. The "Drop ×" column represents the relative reduction in fit when all epistatic terms are constrained. Naturally, the reduction is

Model	p_1, p_2	NCP	Drop ×	Model	p_1, p_2	NCP	Drop ×
1	.5	1.00	.00	7	.5, .5	.55	.02
2	.5	1.06	.00	7	.1, .1	.54	.07
2	.1	1.01	.00	7	.9, .9	1.01	.00
2	.9	1.33	.00	8	.5, .5	.51	.01
3	.5	1.49	.00	8	.1, .1	.66	.00
3	.1	1.03	.00	8	.9, .9	.54	.11
				8	.9, .1	.99	.00
4	.5, .5	.48	.11	9	.5, .5	.48	.11
4	.1, .1	.51	.00	9	.1, .1	.51	.05
4	.9, .9	.45	.48	9	.9, .9	.64	.00
4	.9, .1	1.09	.01	9	.1, .9	.91	.00
5	.5, .5	.69	.06				
5	.1, .1	.91	.00	10	.5, .5	.46	.60
5	.9, .9	.61	.36	10	.1, .1	.49	.05
6	.5, .5	.43	.06	11^1	.5, .5	.52	.13
6	.1, .1	.49	.08	11^2	.5, .5	.43	.35
6	.9, .9	.46	.29	12	.5, .5	.45	.36
6	.1, .9	.78	.01	13	.5, .5	.54	.06

Table 5.5: Full model NCP per sibship (as a proportion of NCP for model M_1 with equal allele frequencies). The column "Drop ×" gives the relative reduction in fit after dropping all epistatic components. For model M_{11}, 11^1 indicates $x = 0.5$ and 11^2 indicates $x = 1$.

0 for the non-epistatic models $M_1 - M_3$. The relative reduction is also small for most other models; models with a lower full model NCP tend to have a greater reduction, representing their greater loading of full model epistatic variance components. Power to detect epistasis, a function of this reduction, is therefore expected to be low.

Not all 13 sets of full results will be tabulated in this section (although all tables can be easily calculated at `http://statgen.iop.kcl.ac.uk/gpc/epistasis.html`). An example table is shown in Table 5.6, for model M_4 with equal allele frequencies. The first ten columns present the QTL and residual variance components under the full model and the various submodels. Two additional columns give the percentage of the full model NCP retained under the submodels (NCP_P) and the percentage of variance under the full model that is also estimated in each submodel (VC_P). For example, if the full model components are 5% $A \times A$ variance and 5% $D \times D$ variance, then VC_P would be 50 for submodels 1 and 2 (which still include the $A \times A$ term),

Model	$A1$	$D1$	$A2$	$D2$	AA	AD	DA	DD	S	N	NCP_P	VC_P
Full	1.3	0.7	1.3	0.7	2.7	1.3	1.3	0.7	20	70	100	100
1	1.5	0.5	1.5	0.5	2.0	2.0	2.0		20.0	70.0	100	93
2	-0.1	0.15	-0.1	1.5	6.1				20.5	70.5	99	67
3	3.0	1.5	3.0	1.5					18.9	72.0	89	40
4	4.6		4.6						18.1	72.7	85	27
5	3.0	1.5							20.8	74.7	44	20
6			3.0	1.5					20.8	74.7	44	20
7	4.5								20.5	75.0	42	13
8			4.5						20.5	75.0	42	13
9									22.7	77.3	0	0

Table 5.6: Example results: model M_4 with $p_1 = 0.5 \;\; p_2 = 0.5$.

and 0 for models 3 to 9 (under the full model, the remaining terms do not account for any variance).

For model M_4, the QTL variance seems fairly evenly distributed among the different sources of variance: $A \times A$ variance accounts for 2.7% of the trait variance, main additive effects at both loci each account for 1.3% of the trait variance. Examining submodel 3 (4^{th} row) with all epistatic components dropped (which account for 60% of the QTL variance – $VC_P = 40\%$) there is only a small drop in the NCP ($NC_P = 89\%$), due to the remaining non-epistatic variance components soaking up some of the unmodelled epistatic variance. The additive effects at each locus now are estimated at 3%, compared to 1.3% under the true model. Continuing to the single-locus submodels 7 and 8, the additive variance for these loci is estimated at 4.5%. Therefore, if the sample were large enough, these might be detectable despite the "true", epistatic nature of the QTL architecture. Conversely, there would be very little power to detect the epistatic effect at work here, as the difference in model fit between the epistatic and non-epistatic models is artificially small. If the M_4 allele "2" is common ($q_1 = q_2 = 0.9$) then full model epistatic components are nearly zero, so the single-locus approximation performs excellently. In contrast, if the "2" allele is rare (10%), the majority of the QTL variance will be attributable to $D \times D$ epistasis. In this case, submodel apparent components of variance can be greatly distorted: in

Model	$A1$	$D1$	$A2$	$D2$	AA	AD	DA	DD	S	N	NCP_P	VC_P
Full	0.0	0.0	0.0	0.0	5.0	0.0	0.0	5.0	20.0	70.0	100	100
1	1.3	-1.3	1.3	-1.3	-0.1	5.1	5.1		19.7	70.3	98	50
2	-2.6	1.3	-2.6	1.3	10.2				21.0	71.5	91	50
3	2.5	1.3	2.5	1.3					18.4	74.0	64	0
4	3.8		3.8						17.7	74.6	61	0
5	2.5	1.3							20.0	76.3	31	0
6			2.5	1.3					20.0	76.3	31	0
7	3.8								19.7	76.6	30	0
8			3.8						19.7	76.6	30	0
9									21.6	78.4	0	0

Table 5.7: Model M_{12} : $p_1 = 0.50$ $p_2 = 0.50$.

submodel 2 both single-locus additive variances are -4.5%, whilst the $A \times A$ variance component is 10% of trait variance.

Now we consider the more extreme epistatic model M_{12} proposed by Frankel and Schork (1996). As shown in Table 5.7, there is evidence for linkage: although the marginal genotypic means are all equal, the expected correlations conditional on joint IBD sharing are not all equal. Rather, the **R** matrix in this case is

$$\begin{pmatrix} 0.20 & 0.20 & 0.20 \\ 0.20 & 0.21 & 0.23 \\ 0.20 & 0.23 & 0.30 \end{pmatrix}$$

given that both loci jointly account for 10% of the trait variance. The single-locus models recover 30% of the information for linkage, estimating the additive QTL variance at just under 4%.

The other extreme epistatic models demonstrate a similar pattern of results: loci are in principle detectable by means of single locus analysis, even if there are no main effects of single loci. For model M_{10}, the checkerboard model, there is not any variance attributable to single-locus components under the full model: nonetheless, individual loci are still detectable using single-locus analysis.

5.4.1 Tiwari and Elston (1998) results reconsidered

For a number of different epistatic models, Tiwari and Elston (1998) examined the associated variance components as a function of the allele frequencies of the trait loci. The premise for this work was, as mentioned above, that the relative magnitude of main versus epistatic effects will directly impact on the utility of single locus versus multi-locus approaches. The authors do not, however, evaluate the actual behaviour of the two-locus Haseman-Elston model in this paper.

As in the present work, the authors restrict themselves to two-locus models with 0/1 means. We shall consider five of these models, which correspond to the current models M_4 through M_8. In the case of a binary trait, the allele frequency at one locus will be fixed if the prevalence, penetrance matrix and allele frequency at the other locus are specified. Tiwari and Elston (1998) assume a population prevalence, $k = 0.1$, and full penetrance, $f = 1$, and give formula for the frequency of the "1" allele at the second locus, p_2 for the different models. For example, for the model corresponding to M_4

$$p_2 = 1 - \frac{\sqrt{k/f}}{1 - p_1}$$

In this way, by only altering a single parameter, p_1, the variance components under different models can be sensibly plotted and compared. Otherwise, altering allele frequencies without this constraint process would result in different implied prevalences, and so the comparison of variance components across the range of allele frequencies would not be valid.

In general, Tiwari and Elston (1998) conclude that for the majority of epistatic models they considered, the epistatic variance components could be greater than the main effects of either of the two individual loci. Although the population prevalence and the specific model of epistasis alter the shapes of variance components curves plotted against allele frequency at one locus, overall this pattern holds. The rarer the

disease, the more likely for epistatic variance components to be the highest. From this they conclude that if a trait is caused by multiple loci and demonstrates epistasis, then single-locus linkage approaches are unlikely to have any power, heralding the need for two-locus linkage models to be more routinely applied.

Figure 5.1 extends these results by reproducing the variance components under the full model and a non-epistatic submodel; in addition, the expected NCP per sib pair under the full and submodel are also plotted. As can be seen, the NCP under both single- and two-locus models primarily refletcs the full model proportion of additive QTL variance. That is, loci with predominantly epistatic effects will be harder to detect whether or not multi-locus methods are applied. Conversely, there is no significant loss in power from only considering single-locus models, even when the model is epistatic. The middle row of figures (the expected variance components under the non-epistatic submodel) show how these are inflated relative to the full model – in all cases, the total QTL variance is still nearly 10% (the true value). This kind of pattern holds for all other models tested.

5.5 Summary

In QTL linkage analysis, additive main effects and epistatic interaction effects are partially confounded because the allele-sharing variables that index epistatic and non-epistatic effects are correlated: e.g. π_1 is correlated with $\pi_1\pi_2$. As a result, variance components under submodels are distorted, with two main implications. First, the analysis of a single locus can in fact detect a QTL with no main effect that interacts epistatically with another (unmeasured) locus. This indicates that single-locus approximations may well be adequate even for the most extreme cases of epistasis, contrary to the warnings of Frankel and Schork (1996). Second, because the apparent variance components in submodels soak up a large proportion of variance attributable

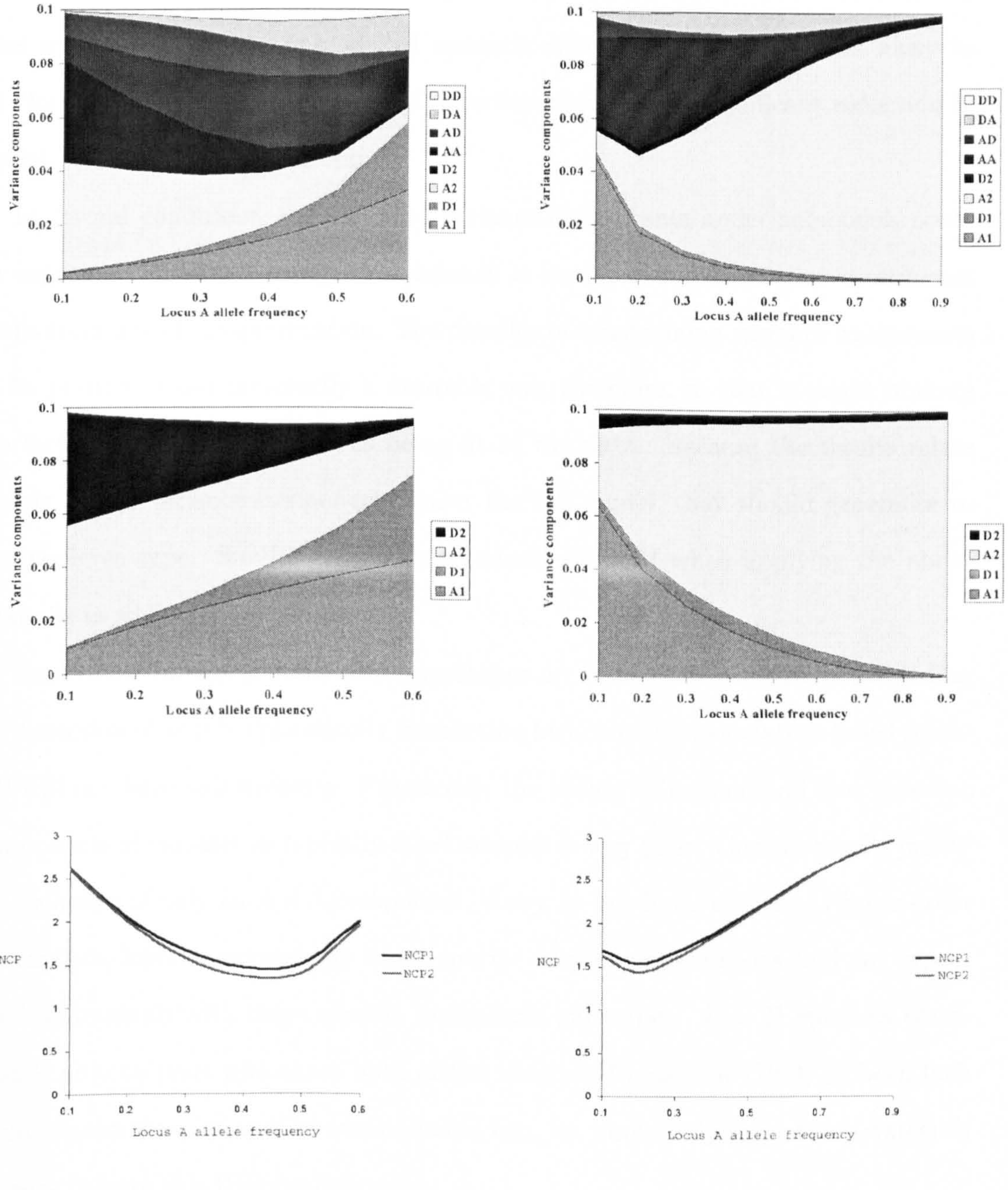


Figure 5.1: Variance components under full model and submodel 4 for M_4 (left column) and M_8 (right column). The top row of figures represent the full model expected variance components. The middle row of figures represents the expected variance components under a non-epistatic submodel: note how these are inflated such that in nearly all cases the total QTL variance still accounts for almost 10% of the trait variance. The bottom row plots the expected NCP, for full model (NCP1) and non-epistatic submodel (NCP2).

to epistatic effects under the full model, power to detect epistasis formally is low. In other words, epistasis will be "informally" detected (when estimates of additive variance are inflated due to unmodelled epistatic effects) although it is less likely to be "formally" detected (where formal detection represents a significant reduction in model fit when an effect is dropped).

In several conditions, the expected variance components under submodels could be negative. Although a negative variance is meaningless, the negative estimates result from model misspecification. The practice of constraining variance components to be positive is not necessarily a desirable one, therefore, in that it might obscure the fact that the wrong model is being fit to the data. Because the results relate mainly to the variance components under the full model, they should generalise to any pedigree type. Similar results were indeed obtained when applying the above methods to sibling trios.

Traditional linkage genome scan approaches are in fact not necessarily precluding the detection of purely epistatically interacting loci, whereas association-based methods will not have this property. Practically, the ability to estimate all four epistatic components of variance in typically-sized samples is very poor. Current results justify the inclusion of only an $A \times A$ component, if any, in two-locus analysis. Higher-order interactions involve increasingly more specific allelic configurations and we expect fewer individuals with these specific multi-locus genotypes. $D \times D$ epistasis corresponds only to pairs who share both alleles identical by descent (IBD) at both loci: in unselected samples and for two-unlinked loci, we would only expect one-sixteenth of pairs to have this IBD configuration.

Chapter 6

Population stratification

This Chapter presents a method for detecting population stratification in samples of unrelated individuals for whom a number of unlinked loci have been genotyped. A latent class analysis model is applied, in which each latent class corresponds to a population stratum. Within strata, Hardy-Weinberg and linkage equilibrium are assumed; for the entire sample, the presence of Hardy-Weinberg and/or linkage disequilibrium across the unlinked loci is indicative of population substructure. For a pre-specified number of hypothetical population classes, the method assigns to each individual the probability of belonging to each class, which may be used as covariates in tests of association to control for stratification effects. Various extensions to the basic model are described, including the ability to model admixture. The method is implemented in the software L-POP *and applied to both real and simulated data.*

6.1 Background

As mentioned in the Introduction and Chapter 3, related individuals can be used to control for population stratification effects, as they are necessarily matched for population strata. An alternative approach is to obtain some index of subpopulation membership, if more than one subpopulation indeed exists within a sample. For

example, self-reported ethnicity could be used as a covariate in tests of association, thereby controlling for stratification. However, there is evidence to suggest that ethnic labels are often inaccurate with regards to underlying genetic differences (Wilson et al., 2001). Furthermore, stratification effects may be subtle and occur within self-reported race. However, this is currently a controversial issue (see the discussion of Risch et al. (2002) at the end of this Chapter).

Another possibility is to use individuals' genetic backgrounds to infer the presence of population substructure. Individuals can then be classified according to the estimated population substructure and tests of association can then take this confounding factor into account. That is, stratification is only corrected for if it is detected in the first instance.

Two approaches that utilise a sample's genetic background to detect and correct for stratification have been suggested, now labelled "genomic control" (e.g. Devlin and Roeder (1999)) and "structured association" (e.g. Pritchard et al. (2000), Pritchard and Donnelly (2001)). Both approaches require multilocus genotype data from across the genome for each individual in the sample. The essence of the genomic control approach is that population stratification leads to a systematic "over-dispersion" of χ^2 statistics in the disease-gene association test. Pritchard and Rosenberg (1999) proposed a test to assess whether or not the χ^2 statistics for a collection of unlinked marker loci across the genome are actually distributed as a χ^2 statistic (i.e. and not over-dispersed): if they are, then the researcher need not worry about stratification. Devlin and Roeder (1999) extended this approach to provide a correction factor for tests of association if in fact stratification was detected. Under no stratification, the test statistics T_N at null, unlinked loci are distributed χ^2_1, whereas in the presence of stratification $T_N/\lambda \sim \chi^2_1$. Devlin and Roeder (1999) developed a method to estimate the multiplicative inflation factor λ which can be used to adjust the statistic at the test locus, T (i.e $T/\lambda \sim \chi^2_1$).

The approximation $\lambda \approx 1 + RF_{ST} \sum_k (f_k - g_k)^2$ expresses the expected inflation factor in terms of sample size (R is number of cases = number of controls), genetic distance between subpopulations (Wright's F_{ST} index) and the proportion of cases and controls from subpopulation k (f_k and g_k respectively). For example, for two equifrequent subpopulations with $F_{ST} = 0.01$ for which a disease is twice as common in one subpopulation, and for a sample consisting of 1000 cases and 1000 controls, then $\lambda \approx 1.5$. In other words, under these conditions, the test statistic would be 150% the size it should be. As association studies utilise larger samples in order to detect genes of very small effect, then the consequences of stratification will also be proportionally magnified – the samples will have more power to detect stratification as a false positive association. Therefore, even if stratification effects are relatively subtle, they may still pose a real danger in modern case-control designs.

Rather than simply estimating a single inflation factor, the structured association approach attempts to assign individuals to subpopulations and to test for association conditional on subpopulation membership. There are two main classes of structured association approaches, which map onto the two main classes of cluster analysis: distance-based and model-based methods. Distance-based approaches (e.g. Schork et al., 2001) proceed by taking some measure of genetic distance (e.g. Nei (1987); Cavalli-Sforza and Edwards (1967); Bowcock et al. (1994)) and some clustering algorithm (e.g. complete linkage, single linkage, centroid clustering) to find clusters in the data. This approach is perhaps complicated by the large number of combinations of distance measures and clustering algorithms that could be employed. Also, the methods do not provide statistical tests to determine whether or not a given solution provides a better or worse fit to the data than any other.

In contrast, model-based clustering methods allow different solutions to be compared statistically. A number of methods have recently been published on model-based approaches to structured association. Pritchard et al. (2000) developed the STRUCTURE

program based on a Bayesian framework and Satten et al. (2001) adopted a latent class analysis (LCA) approach within a maximum-likelihood (ML) framework, using the E-M algorithm. The present work also adopts a LCA approach similar to Satten et al. (2001), albeit with various extensions. Although Bayesian and ML approaches differ in the statistical apparatus employed, they share the same underlying model, which will be described below.

Structured association offers certain advantages over the genomic control approach. First, any structure in a sample is of intrinsic interest – rather than simply computing a single inflation factor, it is far more informative to classify individuals into meaningful groups. Structured association can handle multi-allelic markers whilst current genomic methods are limited to diallelic markers. Structured association can also handle allelic heterogeneity between subpopulations – subpopulation membership can be entered as an interaction term as well as a covariate in any subsequent association test. Finally, structured association does not assume that the genetic distance between two groups is constant across the genome, unlike genomic controls methods.

6.2 Method

A population is assumed to consist of K hidden sub-populations. The basic model assumes that each individual belongs to one and only one sub-population, that mating occurs randomly within each sub-population and that these sub-populations may vary in allele frequencies at loci all across the genome. The aim is to breakdown a population that, as a whole, potentially displays Hardy-Weinberg and linkage disequilibrium across unlinked loci into a number of sub-populations, so that within each sub-population there is Hardy-Weinberg and linkage equilibrium. Consider a number of unlinked genotyped markers across the genome: in a non-stratified sample, one would not expect to observe correlations between these loci (i.e they should be in

linkage equilibrium). In a stratified sample, one would not expect to observe correlations between these loci *within subpopulation.* In practice, the markers do not necessarily need to be completely unlinked: they must be sufficiently distant to be in linkage equilibrium within subpopulation (about 1 cM in homogeneous populations).

The approach is implemented using a latent class analysis model to calculate the best-fit number of latent classes (i.e. $K > 1$ is evidence of stratification). For each individual, the posterior probabilities of belonging to each latent class conditional on genotype data are calculated: these quantities can be used as covariates to control for stratification in subsequent association analyses. This model is similar to that used by Satten et al. (2001), who also employed some additional statistical methods to improve starting values and select solutions. The novel extensions presented in this Chapter involve the inclusion of admixture models; additionally, more extensive simulation results and consideration of various implementation issues are presented.

6.2.1 Latent class analysis

The aim of latent class analysis (LCA, Lazarsfeld and Henry (1968)) is to probabilistically assign individuals to class C of K possible classes on the basis of their responses to multiple variables. In the present context, each class C corresponds to a potential population stratum; individuals' responses correspond to sets of genotypes measured on unlinked loci, G.

The LCA model involves three inter-related sets of probability values $P(C|G)$, $P(G|C)$ and $P(C)$. For a specific K, the main values to be estimated are the posterior class probabilities $P(C|G)$: the probability that an individual belongs to a subpopulation conditional on genotypic configuration. The E-M algorithm (Dempster et al., 1977) is used to iteratively calculate $P(C|G)$ by estimating $P(G|C)$ and $P(C)$. In this context, $P(G|C)$ represents class-specific allele frequencies - the probability that an individual picked from a certain class has a certain allele at a particular locus.

The other set of parameters, $P(C)$, are the prior probabilities of class membership: the probability that an individual picked at random belongs to class C irrespective of G. For $K > 1$, $P(C)$ represents the mixing proportions of the various classes. Critically, $P(C|G)$ are calculated under the assumptions of Hardy-Weinberg and linkage equilibrium holding *within* each class.

For a given set of $P(C|G)$, $P(G|C)$ and $P(C)$ the likelihood of the sample can easily be calculated; the likelihood is used both to determine convergence of the E-M algorithm and to assist in the comparison of solutions with different K. Evidence for stratification corresponds to the best-fitting solution having $K > 1$. In practice, the parameter sets and the corresponding likelihood are assessed for all models from $K = 1$ to $K = K_{max}$ where K_{max} may be based on prior knowledge of the sample and the quality of the data (sample size, number of loci, etc) but typically has values of around 5 or 6. Within reasonable limits discussed below, this method applies equally to any number of multi-allelic loci.

6.2.2 E-M algorithm

For a sample of N unrelated individuals, the aim is to probabilistically assign each individual i to each of K classes. The posterior probability of individual i belonging to class j is $P(C = j|G_i)$. The relative frequency in class j of allele k at locus l is $P(G_l = k|C = j)$. Note that when G is indexed by an i subscript, it refers to an individual's multilocus genotype; when G is indexed by l, it refers to a single locus in the entire population.

The E-M algorithm proceeds in two steps; the expectation, or E-step, involves calculating the values of $P(C)$ and $P(G|C)$ implied by $P(C|G)$; the maximisation, or M-step, involves recalculating $P(C|G)$ given the new estimates of $P(C)$ and $P(G|C)$. These two steps then iterate until convergence. Prior to iterating the E-M algorithm, initial starting values for $P(C = j|G_i)$ are randomly generated with the constraints

that $P(C = j|G_i) \leq \frac{1}{K-1}$ for $j = 1$ to $j = K - 1$ and $\sum_j P(C = j|G_i) = 1$. Additionally, class-specific allele frequencies are all set at sample allele frequency values. More sophisticated starting value schemes are possible, however, and will be discussed below.

E-Step : $P(C)$ and $P(G|C)$

On any one algorithm iteration, the count $\mathcal{I}$ of individuals in class j of K is obtained by summing over all i individuals

$$\mathcal{I}(C = j) = \sum_i P(C = j|G_i).$$

Note that for $0 < P(C|G) < 1$ individuals are *probabilistically* assigned to classes (i.e. counting in fractions of individuals). The allele counts $\mathcal{A}$ for each class are calculated in an analogous fashion

$$\mathcal{A}(G_l = k|C = j) = \sum_i P(C = j|G_i)\,(\mathcal{D}_{i1} + \mathcal{D}_{i2})$$

where, for nonmissing allele data, $\mathcal{D}_{i1}$ is 1 if individual i's first allele at locus l is k and otherwise 0; $\mathcal{D}_{i2}$ is similarly defined for the individual's second allele. For missing data at a locus, values are imputed into $\mathcal{D}_{i1}$ and $\mathcal{D}_{i2}$ for each possible allele k to represent the probability of that allele occurring in that individual, which equals $P(C = j|G_i)P(G_l = k|C = j)$ where $P(G_l = k|C = j)$ is the estimated class-specific allele frequency from the previous E-M iteration (or the starting values on the first iteration).

Having counted the number of individuals in each class $\mathcal{I}(C = j)$ and the number of alleles in each class $\mathcal{A}(G_l = k|C = j)$, the revised prior class probabilities are simply

$$P(C = j) = \frac{\mathcal{I}(C = j)}{N}$$

whilst revised class-specific allele frequencies are

$$P(G_l = k|C = j) = \frac{\mathcal{A}(G_l = k|C = j)}{2\mathcal{I}(C = j)}$$

as class j contains $2\mathcal{I}(C = j)$ chromosomes.

M-Step : $P(C|G)$

In estimating $P(C|G)$, the probability of observing individual i is first calculated

$$P(G_i) = \sum_j P(C = j) \prod_l \tau P(G_l = k_{i1}|C = j) P(G_l = k_{i2}|C = j)$$

where k_{i1} and k_{i2} are the two alleles at locus l and $\tau = 1$ if $k_{i1} = k_{i2}$ (i.e. homozygous genotype) or $\tau = 2$ if $k_{i1} \neq k_{i2}$ (i.e. heterozygous genotype). To handle missing data, $P(G_l = missing|C = j)$ is defined as 1 and so will not contribute to the product term. It is this step that defines the intra-class properties of Hardy-Weinberg and linkage equilibrium: within each subpopulation all alleles are assumed to occur independently within and across loci. Therefore, the class-conditional probability of observing an individual is simply the product of the relevant class-specific allele frequencies.

Summing over all classes weighted by the prior class probability then gives the overall likelihood of observing that individual, $P(G_i)$. Bayes Theorem is applied to give the posterior probabilities, in the form

$$P(C|G) = \frac{P(G|C)P(C)}{\sum_j P(G|C)P(C)}.$$

Therefore, for individual i the posterior probability of belonging to class j is

$$P(C = j|G_i) = \frac{P(C = j) \prod_l \tau P(G_l = k_{i1}|C = j) P(G_l = k_{i2}|C = j)}{\sum_{j'} P(C = j') \prod_l \tau P(G_l = k_{i1}|C = j') P(G_l = k_{i2}|C = j')}$$

whilst the sample log-likelihood on E-M iteration n is $\lambda_n = \sum_i \ln P(G_i)$. The E-M

algorithm converges if $|\lambda_n - \lambda_{n-1}|$ falls below some arbitrary tolerance value. Otherwise, returning to the E-Step, $P(G|C)$ and $P(C)$ are recounted on the basis of the newly-revised estimates of $P(C|G)$.

6.2.3 AIC model fit criterion

As well as estimating $P(C|G)$ for $K = 1, 2, ...$ one wants to ask: does a more complex model (i.e. higher K) provide a significantly better description of the data? In particular, is there evidence of *any* stratification (i.e. $K > 1$)? As different solutions involve different numbers of unique parameters and are not nested, the Akaike Information Criterion (AIC) (Akaike, 1974), defined as minus twice the log-likelihood plus twice the number of model parameters, is used to evaluate different models.

There are $K-1$ non-redundant parameters in $P(C)$ and $\sum_l (K(n_l - 1))$ in $P(G|C)$ if locus l has n_l alleles. (Posterior class probabilities $P(C|G)$ do not count as separate parameters, as these are implied from $P(C)$ and $P(G|C)$.) The lowest AIC solution is the most parsimonious and best-fitting explanation of the data. In the absence of any *a priori* considerations regarding population substructure, only the $P(C|G)$ from the K-solution with the lowest AIC should be used as covariates in any subsequent association analysis.

6.2.4 Correction for stratification

Whereas the approach of Satten et al. (2001) combines the test of association for binary disease traits with the detection of stratification, the current approach separates these two aspects of the problem. The simple strategy advocated here is to use posterior probabilities $P(C = 1|G)$ to $P(C = K - 1|G)$ from the best-fit solution as covariates in whatever test of association is required. Alternatively, individuals can be assigned to discrete classes on the basis of their highest $P(C|G)$ (although this can induce a bias if the highest posterior probabilities are not very near 1). In

this way, there is no constraint on the type of association test used – the approach is applicable to any type of analysis or trait. As mentioned above, it is also easy to specify interactions by subpopulation to allow for allelic heterogeneity.

Whether or not this approach is optimal, as well as the effect of using covariates derived from a model where K has been over- or under-estimated, will be explored in the future: the rest of this Chapter is concerned with detection of stratification rather than correction. Chapter 9 presents two approaches to testing for association that use $P(C|G)$ to correct for potential stratification.

6.2.5 Admixture models

So far we have assumed a simple population genetic model: K distinct subpopulations of varying size that differ in allele frequencies at unlinked markers; also that Hardy-Weinberg and linkage equilibrium exist within each subpopulation. A more general and realistic model allows for admixture between subpopulations. That is, we may wish to characterise as *admixed* individuals who have descended from two or more other subpopulations also seen in the sample, rather than assuming that a further distinct class exists. Such a model is potentially more powerful and more revealing of hidden population structure.

Admixture is modelled in terms of a finite number of *derived classes* (C_D) that represent an admixture of one or more *ancestral classes* (C_A). Considering discrete sets of admixture proportions by constraining possible proportions to a $1/r$ resolution, we can enumerate all possible derived classes for a given number of ancestral classes. For example, if $r = 2$ and there are 3 ancestral classes, six derived classes are implied.

The matrix

$$\Theta = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.50 & 0.00 & 0.50 \\ 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 1.00 \end{bmatrix}$$

represents the mixing proportions of the three ancestral classes (columns) in the six derived classes (rows). Three of these derived classes are *pure* in the sense that they are derived from only one ancestral class, the other three derived classes are *admixed.* Reading across rows, the elements of Θ represent the proportion of an individual's genome that is derived from each ancestral class.

E-step : $P(G|C_D)$ and $P(C_D)$

Counting individuals, rather than alleles, is still straightforward: individuals are counted directly into derived classes. Unlike the allele counts, there are no constraints on the expected individual counts of the derived classes in terms of the frequencies of the ancestral classes. The prior derived class probabilities are therefore simply estimated as $P(C_D = d) = \mathcal{I}(C_D = d)/N$ where $\mathcal{I}(C_D = d) = \sum_i P(C_D = d|G_i)$.

Rather than directly counting alleles into derived classes, the two layers of classes must now be considered. Of primary interest are the parameters for the derived classes, which correspond to the simple classes considered previously: posterior probabilities are only calculated for the derived classes, $P(C_D|G)$. The presence of the ancestral classes effectively places constraints on how the allele-counting step proceeds, however. Alleles are counted first into ancestral classes, from which the derived class counts are calculated. If a derived class is a 50:50 admixture of two ancestral classes, the derived class allele frequency will be related to the allele frequencies of the two ancestral classes, in this case it will be the average. As such, we effectively reduce

the number of parameters needed whilst constraining the pattern of allele frequencies between derived classes to model admixture.

In the no-admixture case, fractions of alleles were counted into classes in proportion to each individual's $P(C|G)$. For example, if for $K = 2$, individual i with alleles k_1 and k_2 at locus l, has $P(C = 1|G_i) = 0.75$ and $P(C = 2|G_i) = 1 - 0.75 = 0.25$, then $\mathcal{A}(G_l = k_1|C = 1)$ would incremented by 0.75, $\mathcal{A}(G_l = k_2|C = 1)$ by 0.75, $\mathcal{A}(G_l = k_1|C = 2)$ by 0.25 and $\mathcal{A}(G_l = k_2|C = 2)$ by 0.25. However, we cannot count alleles into ancestral classes in the same manner for the following reason. For a model with two ancestral classes and three derived classes (shown in Figure 6.1) consider that for individual i with alleles k_1 and k_2 at locus l, $P(C_D = 2|G_i) = 0.50$ where $C_D = 2$ represents the admixed derived class.

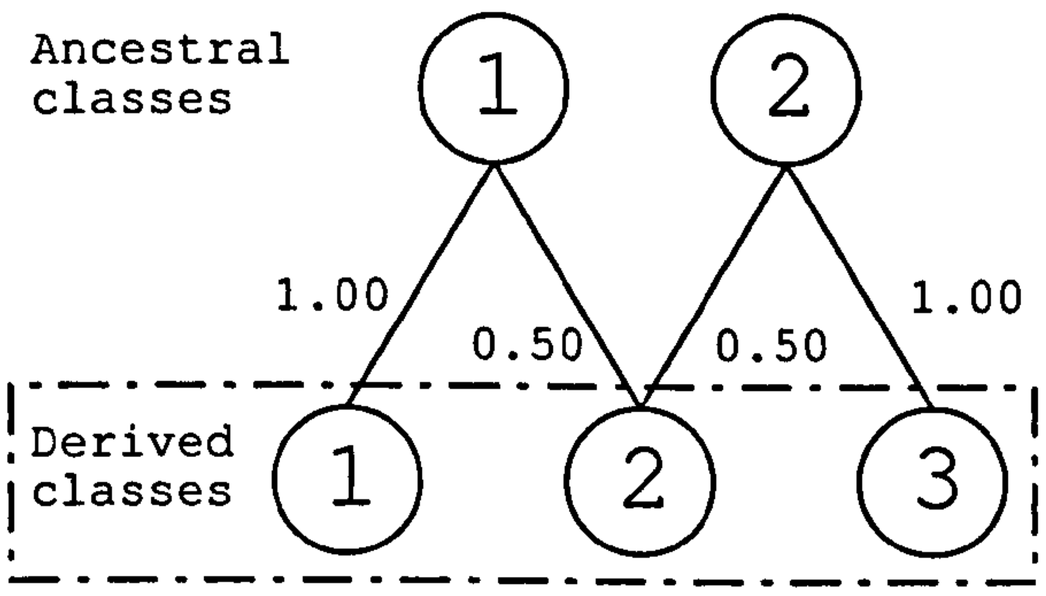


Figure 6.1: Ancestral and derived classes.

Considering allele k_1, we can not simply increment the derived class count $\mathcal{A}(G_l = k_1|C_D = 2)$ by 0.5 and therefore the ancestral class counts $\mathcal{A}(G_l = k_1|C_A = 1)$ and $\mathcal{A}(G_l = k_1|C_A = 2)$ by 0.25 each. A k_1 allele belonging to derived class $C_D = 2$ does not necessarily have an equal probability of having derived from $C_A = 1$ as opposed to $C_A = 2$. We therefore cannot simply assign fractions of alleles to ancestral classes from a derived class in proportion to the known mixing frequencies in Θ. In fact, the probability of ancestral origin will depend upon the allele frequencies of the ancestral classes. If $P(G_l = k_1|C_A = 1)$ is 0.1 whilst $P(G_l = k_1|C_A = 2)$ is 0.2, it is twice as

likely that the allele k_1 fractionally counted into derived class $C_D = 2$ originated from ancestral class $C_A = 2$, given that the mixing proportions of $C_A = 1$ and $C_A = 2$ are equal in $C_D = 2$. Of course, we are counting alleles into ancestral classes precisely in order to calculate ancestral class allele frequencies: it is sufficient to use the ancestral class allele frequencies from the previous E-M iteration as these will converge to the true estimates.

In general, we calculate the expected contribution from ancestral class a of allele k conditional on the data and current model estimates as the unit to be used in the ancestral class allele count. For individual i, considering allele k at locus l, we can calculate the expected contribution from ancestral class a

$$\alpha(G_l = k|C_A = a, G_i) = \sum_d P(C_D = d|G_i) \left[\frac{[\Theta]_{da} P(G_l = k|C_A = a)}{\sum_{a'} [\Theta]_{da'} P(G_l = k|C_A = a')} \right]$$

where the first sum is over d derived classes. (Note that if $\sum_{a'} [\Theta]_{da'} P(G_l = k|C_A = a')$ equals zero, it can be set to any nonzero number without adverse effect, to avoid computational problems.)

The sample contribution to the allele k count for ancestral class a is therefore

$$\mathcal{A}(G_l = k|C_A = a) = \sum_i \alpha(G_l = k|C_A = a, G_i) (\mathcal{D}_{i1} + \mathcal{D}_{i2})$$

which is equivalent to the original allele-counting formula in the case of a "pure" derived class where $[\Theta]_{da}$ takes only 1 or 0 values as $\alpha(G_l = k|C_A = a)$ will only ever equal $P(C_D = d|G_i)$ or 0.

Missing data have to be handled slightly differently, however. The contribution to the ancestral class count a for each possible allele k, is calculated by summing over all derived classes d

$$\mathcal{A}(G_l = k|C_A = a) = \sum_i \left(\sum_d \alpha(G_l = k|C_A = a, G_i) [\Theta]_{da} \right) (\mathcal{D}_{i1} + \mathcal{D}_{i2})$$

where, as before, for missing alleles $\mathcal{D}_{i1} = \mathcal{D}_{i2} = P(C_D = d|G_i)P(G_l = k|C_D = d)$.

The ancestral class individual counts can be calculated by simply summing over all the k allele counts for any one locus l

$$\mathcal{I}(C_A = a) = \sum_i \sum_k \mathcal{A}(G_l = k|C_A = a)/2$$

Having counted the number of individuals and alleles in each ancestral class, we calculate the allele frequencies in the ancestral classes

$$P(G_l = k|C_A = a) = \frac{\mathcal{A}(G_l = k|C_A = a)}{2\mathcal{I}(C_A = a)}$$

and then finally the derived class allele frequencies which are simply weighted sums of the constituent ancestral class allele frequencies

$$P(G_l = k|C_d = d) = \sum_a P(G_l = k|C_A = a)[\Theta]_{da}$$

M-step : $P(C_D|G)$

Having calculated the derived class prior probabilities $P(C_D)$ and allele frequencies $P(G|C_D)$, the M-step proceeds, for derived classes only, as in the no-admixture case described above.

6.2.6 Fixing individual posterior classes probabilities

It may sometimes be desirable to allocate individual i to latent classes j, by fixing $P(C = j|G_i)$ to 1 and $P(C \neq j|G_i)$ to 0, rather than estimating these values. This procedure allows the likelihood to be calculated for any classification of individuals based on external criteria (e.g. self-reported ancestry). Additionally, this procedure can be used to "anchor" the solution: for example, if the sample contains a few "prototypical" individuals (i.e. those with unambiguous ethnic group information)

then these individuals can be fixed to specific classes. This is particularly useful when more complex admixture models are specified.

Although not yet implemented, fixing individuals to classes could also help in setting starting values: if the class-specific allele frequency values were based on the few fixed class exemplars (instead of every class using the sample allele frequencies), this could presumably aid E-M convergence.

6.2.7 Haploid and X chromosome data

Although the method above applies to diploid genotypic data, a straightforward modification enables the analysis of haploid organisms, or of X chromosome data in males. In particular, only one allele at each locus is now counted in the E-step

$$\mathcal{A}(G_l = k|C = j) = \sum_i P(C = j|G_i)\mathcal{D}_1$$

and so the class-specific allele frequencies are now

$$P(G_l = k|C = j) = \frac{\mathcal{A}(G_l = k|C = j)}{\mathcal{I}(C = j)}$$

whilst in the M-step, the calculation of $P(C|G)$ becomes

$$P(C = j|G_i) = \frac{P(C = j)\prod_l P(G_l = k_{i1}|C = j)}{\sum_{j'} P(C = j')\prod_l P(G_l = k_{i1}|C = j')}$$

6.2.8 Genetic outlier detection

A related goal to detecting subpopulations within a sample is the detection of population outliers using genetic background information. That is, the sample may be relatively homogeneous except for one or two individuals. These individuals would not constitute a class by themselves – but it might be of interest to identify such

individuals before embarking on any other analyses. A proposed method is first to calculate the sample log-likelihood $\ln L_0$ for $K = 1$. Then, for each individual i, the sample log-likelihood $\ln L_i$ is calculated for $K = 2$ but with individual i fixed to class 2 (i.e. fix $P(C = 2|G_i) = 1$) and all other individuals fixed to class 1 (i.e. fix $P(C = 1|G_m) = 1$ for $m \neq i$). The difference $\ln L_i - \ln L_0$ is a measure of genetic distance and can be inspected to identify genetically outlying individuals. A similar approach has recently been proposed by Fisher and Lewis (2001).

6.2.9 Model diagnostics

In order to aid the model-fitting process, several diagnostic statistics have been implemented in L-POP.

Inter-class genetic distance matrix

An inter-class genetic distance matrix using Nei's measure of genetic distance (Nei, 1987) is calculated from the class-specific allele frequencies. Nei's genetic distance between two classes, j_1 and j_2, for N loci is calculated

$$d_{\text{Nei}} = -\ln \frac{\sum_{l=1}^{N} \sum_k \left[P(G_l = k|C = j_1)P(G_l = k|C = j_2)\right]/N}{\sqrt{\sum_{l=1}^{N} \sum_k \left[P(G_l = k|C = j_1)^2\right] \sum_{l=1}^{N} \sum_k \left[P(G_l = k|C = j_2)^2\right]/N}}$$

It is especially convenient to apply a multidimensional scaling technique to the distance matrix, in order to obtain a visual representation of the class structure. An example of such a plot based on real data is given in the section below dealing with the Wilson *et al* dataset.

Classification entropy

An 'entropy' measure is calculated for each individual, to indicate how well that individual has been classified in the final solution. For example, considering the

following two individuals, clearly ID1 has been classified with more certainty than ID2. . .

Ind	$P(C=1\|G)$	$P(C=2\|G)$	$P(C=3\|G)$	Entropy
ID1	0.01	0.00	0.99	0.056
ID2	0.10	0.30	0.60	0.898

Entropy for individual i is calculated by summing over all j classes 1 to K: $-\sum_{j=1}^{K} P(C=j|G_i) \ln P(C=j|G_i)$ where $P(C=j|G_i) > 0$. The measure ranges between 0 and 1, where a lower value represents a better classification.

Locus-specific distances

Inter-class Nei genetic distances are also calculated for each locus separately. These statistics can be useful for identifying which loci are contributing to solutions with $K > 1$. Typically, one would expect all loci to contribute approximately equally. In cases where only a couple of loci stand out as contributing much more than the others, it is worth investigating the positions of these loci – it might be indicative of the loci being tightly linked. In this case, at least one of the markers should be removed from the dataset. Class-specific locus-specific genetic distances are also calculated (i.e. comparing class j against all other classes for that locus).

6.2.10 Comparing solutions

An auxiliary utility `COMPSOL` can be used to compare different solutions from `L-POP` either against each other or an external classification scheme. For each solution, the data are partitioned by assigning each individual to a single class based on highest posterior probability. A two-way contingency table is constructed for each pair of solutions. Inspection of the contingency tables can be useful to see the hierarchical structure of the cluster solution. For example, comparing $K = 2$ and $K = 3$ solutions,

as shown below, indicates that class "2" in the two-class solution splits into classes "2" and "3" in the three-class solution:

	Three classes		
Two classes	1	2	3
1	50	0	0
2	0	25	25

The adjusted RAND index (Hubert and Arabie, 1985) is a measure of agreement specifically designed to compare partitioning schemes of data from clustering methods. The adjusted RAND index varies between 0 and 1 (where 0 represents no agreement and 1 represents complete agreement) and is calculated

$$\text{RAND} = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{\cdot j}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2}\right] - \left[\sum_i \binom{n_{i\cdot}}{2}\sum_j \binom{n_{j}}{2}\right]/\binom{n}{2}}$$

and is displayed for all pairwise combinations of solutions by `COMPSOL`.

6.2.11 Implementation

The methods described above have been implemented in the computer program `L-POP`[1]. The program can handle missing genotypic data, autosomal and X chromosome markers and haploid organisms. Posterior probabilities can be estimated for each individual; alternatively, individuals can be fixed to belong to a particular class. Options to specify admixed solutions, relax certain assumptions and calculate the diagnostic measures mentioned above are incorporated.

The number of parameters can increase very quickly for "wide" datasets with increasing K however: for just 50 SNP markers and $K = 2$ there are 101 parameters, and 203 parameters for $K = 4$. For many larger problems (with a large number of individuals and/or loci, and especially when K is greater than 2) the maximum-

[1] `L-POP` is available for download from `http://statgen.iop.kcl.ac.uk/lpop/`.

likelihood approach using the E-M algorithm can suffer from computational problems, especially local minima. A feature of `L-POP` is to automatically restart the algorithm with different starting values a user-defined number of times, and to pick the minimum of all the repeated converged solutions. An option which could be added in the future would be a method to choose more sensible starting values – this should aid convergence (see Satten et al. (2001)).

Additionally, a program was designed to easily simulate simple datasets from a number of discrete subpopulations (including admixed classes). In its basic form, the `L-SIM` program requires the number of ancestral classes to be specified, along with different groups of loci with a set number of alleles and ancestral-class-specific allele frequencies. Samples may then be generated, given the number of derived classes required, the number of individuals in each derived class and the mixing proportions (matrix Θ).

6.2.12 A simple example

To illustrate the use of the method, consider the following sample of five individuals with genotypes for five unlinked markers. All the markers are SNPs with alleles coded 1 and 2; missing alleles are coded 0. Clearly, this sample is not in Hardy-Weinberg or linkage equilibrium. Furthermore, if we assume this population in fact consists of a number of subpopulations which are themselves in Hardy-Weinberg and linkage equilibrium, it is clear that the first two individuals and the second two individuals come from two different populations: in the first subpopulation, the "2" allele does not exist, in the second subpopulation, the "1" allele does not exist. We cannot make any predictions on the basis of the fifth individual's all-missing data.

```
ID1  1/1  1/1  1/1  1/1  1/1
ID2  1/1  1/1  1/1  1/1  1/1
ID3  2/2  2/2  2/2  2/2  2/2
```

ID4	2/2	2/2	2/2	2/2	2/2
ID5	0/0	0/0	0/0	0/0	0/0

Tabulating the AIC for solutions $K = 1$ to $K = 3$ in Table 6.1, we see that a two-class solution is favoured as the most parsimonious explanation of the data, with the lowest AIC of 27.55. The prior class probabilities are $P(C = 1) = 0.5$ and $P(C = 2) = 0.5$ for the two classes: the model suggests the sample comprises of two distinct, equifrequent subpopulations.

K	-2LL	AIC	$P(C = 1)$	$P(C = 2)$	$P(C = 3)$
1	55.45	65.45	1.00		
2	5.55	27.55	0.50	0.50	
3	5.55	39.55	0.50	0.28	0.22

Table 6.1: Basic example results: sample log likelihood, AIC and $P(C)$.

Examining $P(C|G)$ for this solution indicates how the individuals have been assigned to classes:

	P(C=1|G)	P(C=2|G)
ID1	0.00	1.00
ID2	0.00	1.00
ID3	1.00	0.00
ID4	1.00	0.00
ID5	0.50	0.50

These values indicate that individuals ID1 and ID2 belong to class "2" with 100% certainty, whereas individuals ID3 and ID4 belong to class "1" also with 100% certainty. The posterior probabilities for ID5 are as expected: in the absence of any information (i.e. all genotype data missing) the posterior class probabilities will equal the prior class probabilities.

To illustrate the admixture model, consider the following example given below. Eye-balling the data would suggest that individuals ID1 and ID2 belong to one class,

ID3 and ID4 belong to a second class and ID5 – ID8 represent an intermediate, admixed version of the previous two classes:

```
ID1   1/1   1/1   1/1   1/1   1/1
ID2   1/1   1/1   1/1   1/1   1/1
ID3   2/2   2/2   2/2   2/2   2/2
ID4   2/2   2/2   2/2   2/2   2/2
ID5   1/1   1/2   2/2   1/2   1/2
ID6   1/2   2/2   1/2   1/1   2/2
ID7   1/2   1/2   1/1   1/2   1/1
ID8   2/2   1/1   1/2   2/2   1/2
```

The best-fit solution does indeed allow for admixture, as shown in Table 6.2. For the final row of the table $K = 2 + 1$ indicates two pure classes with an additional admixed class that is a 50:50 mixture of the first two: this solution has the lowest AIC value. The predictions from the $K = 2 + 1$ solution are identical to the $K = 3$ solution in this case; the $K = 2+1$ solution is preferred on the grounds of parsimony.

K	-2LL	AIC	$P(C = 1)$	$P(C = 2)$	$P(C = 3)$
1	97.041	107.041	1.0000		
2	69.591	91.591	0.6262	0.3738	
3	58.209	92.209	0.2495	0.5010	0.2495
2+1	58.209	82.209	0.2495	0.5010	0.2495

Table 6.2: Example with admixture results: sample log likelihood, AIC and $P(C)$.

The posterior probabilities $P(C|G)$ for the best-fit solution are

```
     P(C=1|G)   P(C=2|G)   P(C=3|G)
ID1  0.998      0.002      0.000
ID2  0.998      0.002      0.000
ID3  0.000      0.002      0.998
ID4  0.000      0.002      0.998
ID5  0.000      1.000      0.000
ID6  0.000      1.000      0.000
```

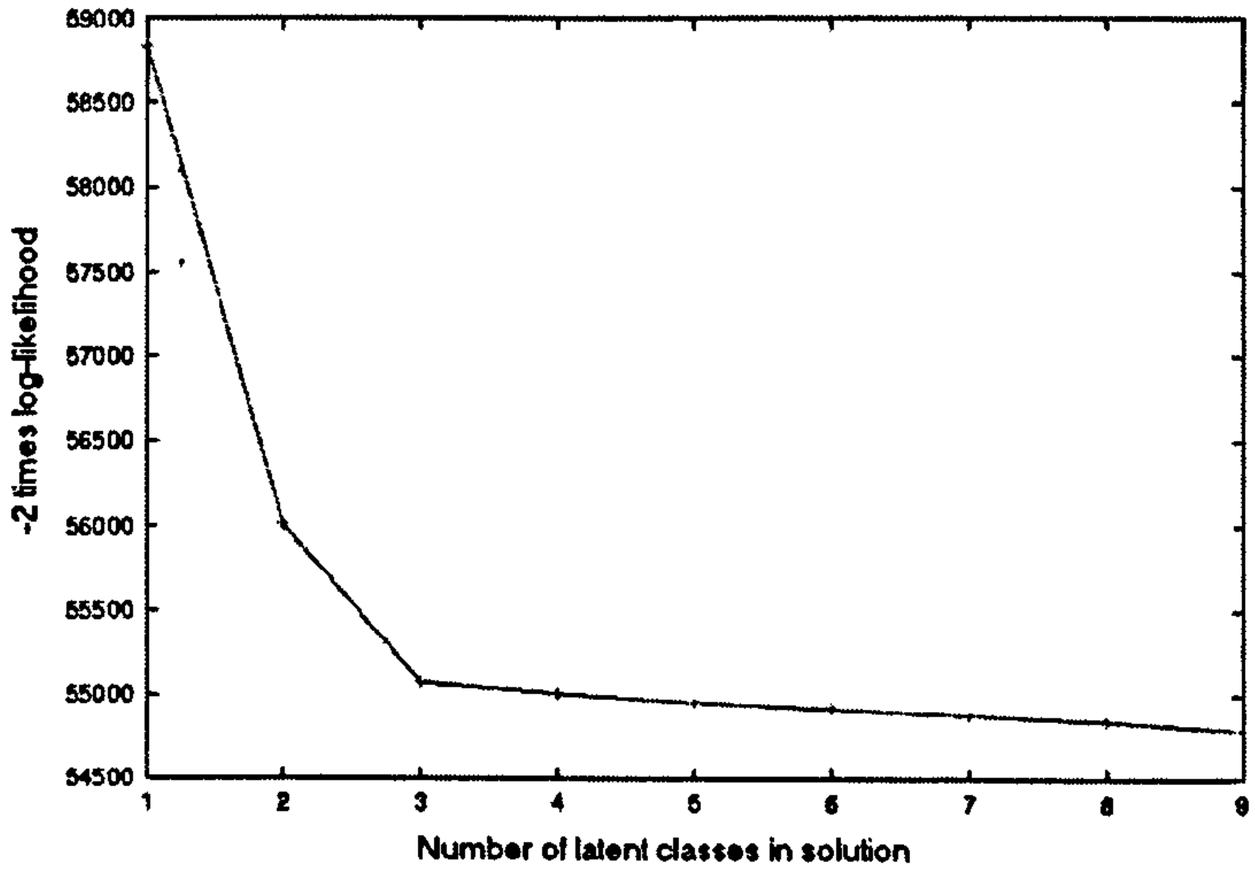


Figure 6.2: Scree-plot of log-likelihood for different LCA solutions: a 3 class solution is optimal.

```
ID7   0.000      1.000      0.000
ID8   0.000      1.000      0.000
```

where the second class is fixed to be the admixed class. The class-specific allele frequencies for one locus allele (they are the same for the other four loci):

```
Class-specific allele frequencies P(G|C)
    l=1     j=1     p(A=1|j)= 1.0000        p(A=2|j)= 0.0000
    l=1     j=2     p(A=1|j)= 0.5000        p(A=2|j)= 0.5000
    l=1     j=3     p(A=1|j)= 0.0000        p(A=2|j)= 1.0000
```

A slightly more realistic example is considered next: 30 SNPs are simulated for 1000 individuals. The individuals were sampled from three subpopulations with mixing frequency 70%, 20% and 10%. Normally-distributed subpopulation-specific random deviations were added to each SNP frequency; the mean of this deviation was 0.0, the variance was 0.2; frequencies were bounded at 0.001 and 0.999. For example, one of the SNPs simulated had frequencies 0.46, 0.36 and 0.54 (frequency of allele "1") in each of the subpopulations; another SNP was 0.2, 0.18, 0.001. The best-fitting solution involved 3 latent classes, thereby mirroring the simulated subpopulation structure. Figure 6.2 shows the log-likelihood under LCA solutions with

different numbers of latent classes (i.e. along the x-axis). Having more than 3 latent classes does not result in a significantly better fit to the data (based on lowest AIC). Under the 3-class solution, the prior probabilities of each latent class were 0.70, 0.09 and 0.21. Using the posterior latent class membership probabilities, only 10 individuals were misclassified out of the 1000, indicating that the method is able to detect this level of population stratification with reasonable accuracy.

6.3 Basic simulations

To further explore the properties of the current method, a number of simulations were conducted. For example, the ability of similar methods (Pritchard et al., 2000) to detect subpopulation structure increase with sample size, number of loci and the degree of divergence between populations. The following simulations are exploratory, in that only a single replicate was generated for each condition. Although it may be desirable to follow-up with more comprehensive simulations which repeat each condition a number of times, the current strategy of numerous 'overlapping' scenarios allows the results to emerge quite clearly.

Thirteen different conditions were examined. In each condition, five datasets were generated, with 10, 20, 50, 100 and 200 marker loci respectively. The different conditions allow the effect of sample size, number of marker loci and genetic distance between subpopulations to be studied as well as other properties of the markers used (e.g. number of alleles and the distribution of between-subpopulation allele frequency differences). In all cases, two models were applied to the data: $K = 1$ and $K = 2$. More complex models are investigated in the 'Further simulations' section.

Original condition

The 'Original' condition simulated 2000 individuals from two subpopulations (P_1 and P_2, with 1000 individuals from each). All marker loci were diallelic, with an average allele frequency of 0.5 and an average between-subpopulation allele frequency difference (δ) of 0.2. For all markers, one allele had a frequency of 0.4 in P_1 and 0.6 in P_2. The results are given in Table 6.3.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	42246.17	41836.74	409.42	0.5574	0.8170	0.314930	0.799972
20	84045.18	82888.07	1157.10	0.4916	0.8935	0.145796	0.837366
50	210651.05	205166.63	5484.41	0.5019	0.9790	0.032647	0.971080
100	421364.24	408004.86	13359.38	0.5021	0.9965	0.006222	0.998072
200	842568.78	813584.40	28984.38	0.5000	1.0000	0.000025	0.999996

Table 6.3: Simulation results: $\delta = 0.2$; $N = 1000 + 1000$; 'Original'.

A two-class solution is correctly favoured in all five conditions (the first column M is the number of marker loci). In the fourth column $\Delta_{AIC} = AIC(K = 1) - AIC(K = 2)$, so a positive value is evidence for a two-class solution over a one-class solution. Even with only 10 markers, this difference is large (409.42). The final four columns refer to parameter estimates under the $K = 2$ solution. The $P(C)$ column gives the prior class probability for class "1" – in all cases this value is near 0.5 (i.e. as the two classes were simulated at equal frequencies), but the estimate increases in precision with increasing number of markers. The *Correct* column gives the proportion of individuals correctly classified on a highest posterior probability basis. For example, comparing true subpopulation membership and estimated class for these hypothetical data:

	Estimated	
True	1	2
1	8	992
2	979	21

we see that estimated class "2" clearly corresponds to true class "1" and vice versa, so the proportion of correctly classified individuals is $(992+979)/(8+992+979+21) = 0.9855$. In the Table 6.3 the classification rate rises from round 80% to 100% as the number of markers increases. That is, although a two-class solution is favoured with only 10 markers, the accuracy of the classification is not perfect. However, for such a small number of markers, arguably 80% accuracy is acceptable.

The final two columns give the average posterior probability for belonging to class "1" for individuals from subpopulation P_1 (7^{th} column) and P_2 (8^{th} column). Perfect classification would correspond to one of these values being 0 and the other being 1. No classification would correspond to both values equalling the prior probability for class "1". (Note that the values have been ordered such that the smaller value always corresponds to P_1 – in practice, whether or not estimated class "1" corresponds to P_1 or P_2 is random and arbitrary.) As can be seen, with increasing number of markers, the separation between the two classes increases – by 100 markers, the classification is almost perfect.

Small sample size

The 'Small' condition was similar to the 'Original' condition, except only 100 individuals from each subpopulation were generated.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	4221.94	4192.87	29.07	0.4851	0.7850	0.250636	0.719524
20	8411.65	8301.43	110.22	0.4833	0.8600	0.151778	0.814751
50	21124.95	20677.30	447.65	0.4877	0.9600	0.034173	0.941307
100	42184.07	40930.80	1253.27	0.5007	1.0000	0.001621	0.999836
200	84448.51	81592.65	2855.87	0.5000	1.0000	0.000000	1.000000

Table 6.4: Simulation results: $\delta = 0.2$; $N = 100 + 100$; 'Small'.

Although a two-class solution is favoured in all cases, the difference in AIC has dropped considerably. However, the accuracy of classification has remained approximately equal to the 'Original' condition. (Note that with the smaller sample size, the precision of the classification estimates themselves will be lower).

Small delta

The 'Delta' condition reduces the genetic distance between the two groups, making them less distinct and therefore harder to separate. In this condition, the δ value is 0.1 instead of 0.2 (i.e all markers are simulated using an allele frequency of 0.45 in the first subpopulation and 0.55 in the second).

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	41562.58	41525.52	37.07	0.6368	0.6485	0.538585	0.735143
20	83300.39	83227.72	72.66	0.4808	0.7205	0.341362	0.620107
50	208380.35	207785.38	594.97	0.5275	0.8475	0.192188	0.752602
100	416906.00	415007.85	1898.15	0.5061	0.9110	0.115835	0.872047
200	834563.87	829149.23	5414.64	0.5000	0.9795	0.030860	0.969244

Table 6.5: Simulation results: $\delta = 0.1$; $N = 1000 + 1000$; 'Delta'.

As Table 6.5 shows, this leads to a reduction in the AIC difference, although a two-class solution is still consistently favoured. The classification ability of the model

also drops under this condition, however. For example, with only 10 markers, the prior probability of class "1" is 0.6368 (i.e. it should be 0.5); the posterior probabilities are both above 0.5 for P_1 and P_2 (i.e one should be near 0, the other near 1). Under these conditions, around 200 markers are required before classification becomes near-perfect.

Small delta and small sample size

The next condition combines the 'Small' and 'Delta' conditions. As shown in Table 6.6, the evidence for the two-class solution is greatly attenuated, especially with a smaller number of markers. With only 10 markers, the model favours a one-class solution, and shows no evidence of classifying individuals correctly (it performs at chance).

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	4242.91	4243.88	-0.97	0.9188	0.5000	0.907926	0.929652
20	8387.60	8384.85	2.75	0.8302	0.5850	0.756527	0.904170
50	20981.54	20965.09	16.46	0.4871	0.8000	0.204908	0.769349
100	41889.19	41805.79	83.40	0.4846	0.8750	0.106652	0.862529
200	83655.70	83297.84	357.87	0.5178	0.9500	0.065752	0.969784

Table 6.6: Simulation results: $\delta = 0.1$; $N = 100 + 100$; 'Delta-Small'.

Summarising the last four conditions, it is clear that small sample size has an extra deleterious effect when conditions are poor to begin with. That is, the small sample size represents 10% of the large sample size. When $\delta = 0.2$, the small-sample Δ_{AIC} is also approximately 10% of the large-sample Δ_{AIC}. For example, for 20 and 200 markers, it is 9.51% and 9.85% respectively (i.e. 110.22/1157.10 and 2855.87/28984.38 respectively). However, when the genetic distance between groups is smaller (i.e. $\delta = 0.1$), then the evidence for stratification is proportionally less in the small sample compared to the large sample: the small-sample Δ_{AIC} is only 3.77% and 6.59% of the

large-sample value, for 20 and 200 markers respectively.

Unequal deltas

However, sample size and average δ value are not the only variables which impact on the model's ability to detect stratification and classify individuals. In the 'Unequal' condition, the subpopulations were simulated with unequal mixing proportions such that one class formed a minority, the other a majority, rather than the 50:50 balance previously used. In this condition, although the overall sample size was held constant (2000), 250 individuals were simulated from the first class, 1750 individuals were simulated from the second.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	41492.66	41357.85	134.81	0.1125	0.911 (0.432)	0.554572	0.935156
20	82927.70	82454.35	473.36	0.1347	0.943 (0.704)	0.308771	0.944791
50	206915.92	204664.96	2250.96	0.1294	0.988 (0.964)	0.012738	0.945832
100	413090.41	407500.63	5589.78	0.1251	0.999 (0.996)	0.000817	0.994949
200	825212.38	813044.21	12168.17	0.1250	1.000 (1.000)	0.000000	0.999948

Table 6.7: Simulation results: $\delta = 0.2$; $N = 250 + 1750$; 'Unequal'.

Compared to the 'Original' condition, there has been some reduction in the Δ_{AIC} values, and the classification ability has been affected also. Two values are given in the *Correct* column for this model – the first is the value calculated as above, the second value in parentheses represents the proportion of the minority subpopulation correctly assigned. For example, for the 10 marker condition, the relationship between true and estimated class was

	Estimated	
True	1	2
1	142	108
2	1713	37

where the overall correct classification is 0.911. Of course, this value is artificially high purely because most individuals will by chance fall in the majority class. However, of the individuals from the minority subpopulation, only 108 of 250 are correctly classified (i.e. estimated class "1" corresponds to the true majority class "2", so only 108 individuals have been placed in a separate estimated class).

Average absolute allele frequency

The next 'Absolute' condition investigated the effect of average allele frequency: keeping δ fixed at 0.2, the allele frequencies were simulated at 0.2 and 0.4 for the two subpopulations rather than 0.4 and 0.6.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	37761.36	37182.45	578.91	0.4776	0.8335	0.209144	0.745869
20	75582.62	73797.11	1785.51	0.4980	0.9150	0.119117	0.876731
50	189000.59	182185.75	6814.84	0.5022	0.9860	0.023420	0.980910
100	378307.88	361713.48	16594.40	0.5001	1.0000	0.000998	0.999170
200	755973.59	720999.45	34974.14	0.5000	1.0000	0.000001	1.000000

Table 6.8: Simulation results: $\delta = 0.2$; $N = 1000 + 1000$; 'Absolute'.

As shown in Table 6.8, there does not appear to be any great effect of absolute allele frequency, compared to the 'Original' condition, at least under these conditions. However, this does not address the issue of whether or not rare alleles are, in practice, more or less likely to show differences between different ethnic groups. Rare alleles are subject to greater fluctuation in frequency due to genetic drift than common alleles, and so may be expected to show greater between-population differences. In fact, recent studies looking at SNP frequencies in different races have concluded that less frequent SNPs are more likely to be specific to one or two races (Cargil et al., 1999; Halushka et al., 1999).

Heterogeneous delta: Split 1

Of course, the average δ value across a set of markers does not capture all the information about allele frequency differences between two groups. In this 'Split1' and the subsequent 'Split2' conditions, the impact of the distribution of frequency differences was examined, whilst keeping the average δ value constant. In the 'Split1' condition, half the markers were simulated to show no difference between groups (i.e. $\delta = 0$, both groups simulated using 0.5 allele frequency) and half the markers were simulated using an exaggerated allele frequency difference ($\delta = 0.4$, groups simulated using 0.7 and 0.3 allele frequencies). In this way, the average between group distance was still $\delta = 0.2$.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	42733.35	41391.74	1341.61	0.5153	0.9035	0.151823	0.878770
20	85379.88	81340.43	4039.44	0.5042	0.9655	0.050230	0.958113
50	213480.61	200146.46	13334.15	0.4998	0.9985	0.002118	0.997496
100	427191.42	397422.33	29769.10	0.5000	1.0000	0.000003	0.999994
200	854385.10	790897.71	63487.39	0.5000	1.0000	0.000000	1.000000

Table 6.9: Simulation results: $\delta = 0.4$, $\delta = 0.0$ (average $\delta = 0.2$); $N = 1000 + 1000$; 'Split1'.

As Table 6.9 shows, the pattern of allele frequency differences gives greater power to detect stratification and better classification rates also. With only 50 markers (25 of which show no between-subpopulation differences) near-perfect classification can be achieved.

Heterogeneous delta: Split2

The 'Split2' condition represents a more extreme version of the 'Split1' condition. Rather than splitting the markers into two equal-sized groups, three-quarters of them were set to show no differences with only the remaining quarter showing an increased δ of 0.8 (i.e. allele frequencies 0.1 and 0.9). (For the 10 marker condition, 2 markers

had $\delta = 0.8$, one marker $\delta = 0.4$ and seven markers $\delta = 0$; similarly for the 50 marker condition). In this way, the average δ value is still 0.2 in all conditions.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	43524.73	39520.60	4004.12	0.4950	0.9860	0.016991	0.973079
20	87332.47	75225.10	12107.36	0.4999	0.9985	0.001512	0.998311
50	218625.59	185358.81	33266.78	0.5000	1.0000	0.000000	1.000000
100	438149.44	367002.43	71147.01	0.5000	1.0000	0.000000	1.000000
200	876360.53	732252.49	144108.03	0.5000	1.0000	0.000000	1.000000

Table 6.10: Simulation results: $\delta = 0.8$, $\delta = 0.0$ (average $\delta = 0.2$); $N = 1000+1000$; 'Split2'.

As Table 6.10 shows, this more extreme split results in even better ability to detect and characterise subpopulation structure, despite the fact that the majority of loci do not show any allele frequency differences between groups at all. This means that a few well-selected markers with large between-group variation might be all that are needed to accurately distinguish between the major ethnic groups.

Multi-allelic markers

All previous simulations have been for a diallelic locus: the method is equally applicable to multi-allelic markers however. A δ value of 0.2 corresponds to $F_{ST} = 0.04$ when the average allele frequency is 0.5. That is, the average expected heterozygosity within each subpopulation is $(1 - 0.6^2 - 0.4^2) + (1 - 0.4^2 - 0.6^2)/2 = 0.48$ and the expected heterozygosity across all populations based on the average allele frequencies is $1 - 0.5^2 - 0.5^2 = 0.50$ and so $F_{ST} = (0.50 - 0.48)/0.50 = 0.04$. In this 'Multi' condition, the performance of using multi-allelic markers with comparable F_{ST} values was examined.

For two populations, the following allele frequency values were used to simulate the markers for the two subpopulations, to give a F_{ST} value of approximately 0.04.

Allele	P_1 frequency	P_2 frequency
1	0.36000	0.14000
2	0.19625	0.30375
3	0.30375	0.19625
4	0.14000	0.36000

As shown in Table 6.11, performance is worse under these conditions, especially for smaller numbers of markers, presumably due to the average between-population allele frequency differences being smaller.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	86594.06	86568.06	26.00	0.5426	0.6520	0.443497	0.641903
20	173601.49	173367.48	234.01	0.4901	0.7630	0.287867	0.692254
50	433520.75	432466.83	1053.92	0.5018	0.8710	0.167834	0.835831
100	866661.64	863705.48	2956.16	0.5016	0.9535	0.068742	0.934548
200	1733146.01	1725047.96	8098.05	0.5005	0.9920	0.012860	0.988230

Table 6.11: Simulation results: Multi-allelic locus, $F_{ST} \approx 0.04$; $N = 1000 + 1000$; 'Multi'.

Multi-allelic markers with rare alleles

In the 'Multi-Absolute' condition, a different set of allele frequencies were employed but with a similar F_{ST} value (0.04). The critical factor in this condition is that some of the subpopulation-specific allele frequencies were set to 0:

Allele	P_1 frequency	P_2 frequency
1	0.2305	0.0000
2	0.3695	0.4000
3	0.4000	0.3695
4	0.0000	0.2305

Table 6.12 shows a markedly different set of results: there is a massive increase in the ability to select a two-class solution and classification is essentially perfect with

only 10 markers. In the 'Multi' condition the average difference in allele frequency between subpopulations was 0.16375; in this 'Multi-Absolute' condition the average difference is even smaller, only 0.1305. It would appear that the presence of allele frequencies of 0 allows the model to easily distinguish between classes.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	89910.59	70606.83	19303.76	0.5000	1.0000	0.000000	1.000000
20	180264.36	138570.70	41693.66	0.5000	1.0000	0.000000	1.000000
50	450743.85	342265.07	108478.78	0.5000	1.0000	0.000000	1.000000
100	901954.54	681723.89	220230.66	0.5000	1.0000	0.000000	1.000000
200	1803227.56	1359667.74	443559.82	0.5000	1.0000	0.000000	1.000000

Table 6.12: Simulation results: Multi-allelic locus, $F_{ST} \approx 0.04$; $N = 1000 + 1000$; 'Multi-Absolute'.

Multi-allelic markers with heterogeneous deltas

In this 'Multi-Split' condition, the three different types of multi-allelic marker used above are combined. Markers with allele frequencies corresponding to the 'Multi-Absolute' condition are labelled 'Type I' markers. Markers with allele frequencies corresponding to the 'Multi' condition are labelled 'Type II'. Finally, markers with four equifrequent alleles (i.e. 0.25 in both subpopulations) are labelled 'Type III' markers.

In all of the five conditions below, only 2 markers are of Type I, 2 are of Type II and the remaining $M-4$ are of Type III. That is, unlike all previous scenarios, where we would expect increasing information with increasing M, only the uninformative marker count rises with M in this condition. In all five cases, there are only four out of M markers that show any difference between subpopulations: when $M = 200$ there are 196 markers which should not contribute anything except noise.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	*Correct*	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	88996.31	86962.74	2033.58	0.5016	0.9450	0.066906	0.936214
20	179469.67	177378.55	2091.13	0.4994	0.9485	0.064293	0.934539
50	449965.40	447953.09	2012.31	0.4960	0.9500	0.052662	0.939306
100	900749.76	898875.57	1874.20	0.5093	0.9465	0.071070	0.947595
200	1802158.95	1800572.82	1586.13	0.4901	0.9485	0.045653	0.934603

Table 6.13: Simulation results: Multi-allelic locus (see text); $N = 1000+1000$; 'Multi-Split'.

The results are shown in Table 6.13: in all cases a two class solution is selected, with large AIC differences, although this decreases with increasing M. The classification ability of the model remains roughly constant over the different M, with a *Correct* value of around 0.95 and posterior probabilities around 0.05 and 0.95. This performance is roughly equivalent to the 'Original' condition with 50 markers – despite the fact that only four markers will be contributing to the solution.

Utilising the diagnostic output features of L-POP, the inter-class locus-specific genetic distances are tabulated for $M = 10$, showing clearly the relative contribution to the solution:

Locus	Type	Inter-class locus variation
1	I	0.6193
2	I	0.6177
3	II	0.0721
4	II	0.0965
5	III	0.0121
6	III	0.0139
7	III	0.0216
8	III	0.0301
9	III	0.0241
10	III	0.0324

Null

The final two conditions examine performance under the null – that is, when there are no true allele frequency differences between subpopulations at any of the markers. Although the two groups are simulated separately, there are no genetic differences between them, so one would expect a single-class solution in all cases. As shown in Table 6.14, this is not necessarily the case, however. In fact, in three out of the five simulations, there was evidence for a two-class solution, although this was quite slight, compared to the previous AIC differences obtained.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	41662.92	41660.40	2.51	0.9894	0.991165	0.987602
20	83211.92	83228.41	-16.49	0.5133	0.506697	0.519974
50	208009.38	208005.18	4.20	0.0453	0.043741	0.046724
100	415331.93	415323.05	8.88	0.0348	0.038148	0.031432
200	831492.11	831513.77	-21.66	0.1422	0.144411	0.139872

Table 6.14: Simulation results: $\delta = 0.0$; $N = 1000 + 1000$; 'Null'.

As expected, the $P(C)$ values (calculated under a two-class solution) are quite meaningless – what is significant is that the P_1 and P_2 posterior probability values are both similar to this value. (Of course, in practice, one would not be aware of the P_1 versus P_2 distinction.) When the model favours a two-class solution, it appears that one of the classes is very small (around 2–4% of the sample).

This suggests that the AIC may have a tendency to over-estimate the true value of K. Further work will be required to investigate the conditions under which the null model is retained, and also the possible use of metrics other than AIC to evaluate model-fit.

Null, small sample size

The final 'Null-Small' condition simulates under the null but with the smaller ($N = 200$) sample size. Results appear to be similar to the 'Null' condition above.

M	AIC(K=1)	AIC(K=2)	Δ_{AIC}	P(C)	$P(C\|G,P_1)$	$P(C\|G,P_2)$
10	4212.41	4211.64	0.77	0.1756	0.174555	0.176442
20	8415.88	8410.18	5.70	0.8743	0.853575	0.895092
50	20908.33	20908.20	0.13	0.8181	0.794096	0.842226
100	41760.40	41791.45	-31.05	0.5517	0.553208	0.550276
200	83470.00	83518.94	-48.94	0.5320	0.537758	0.526349

Table 6.15: Simulation results: $\delta = 0.0$; $N = 100 + 100$; 'Null-Small'.

Summary of simulations

Figure 6.3 plots the AIC difference for the 13 different conditions as a function of M.

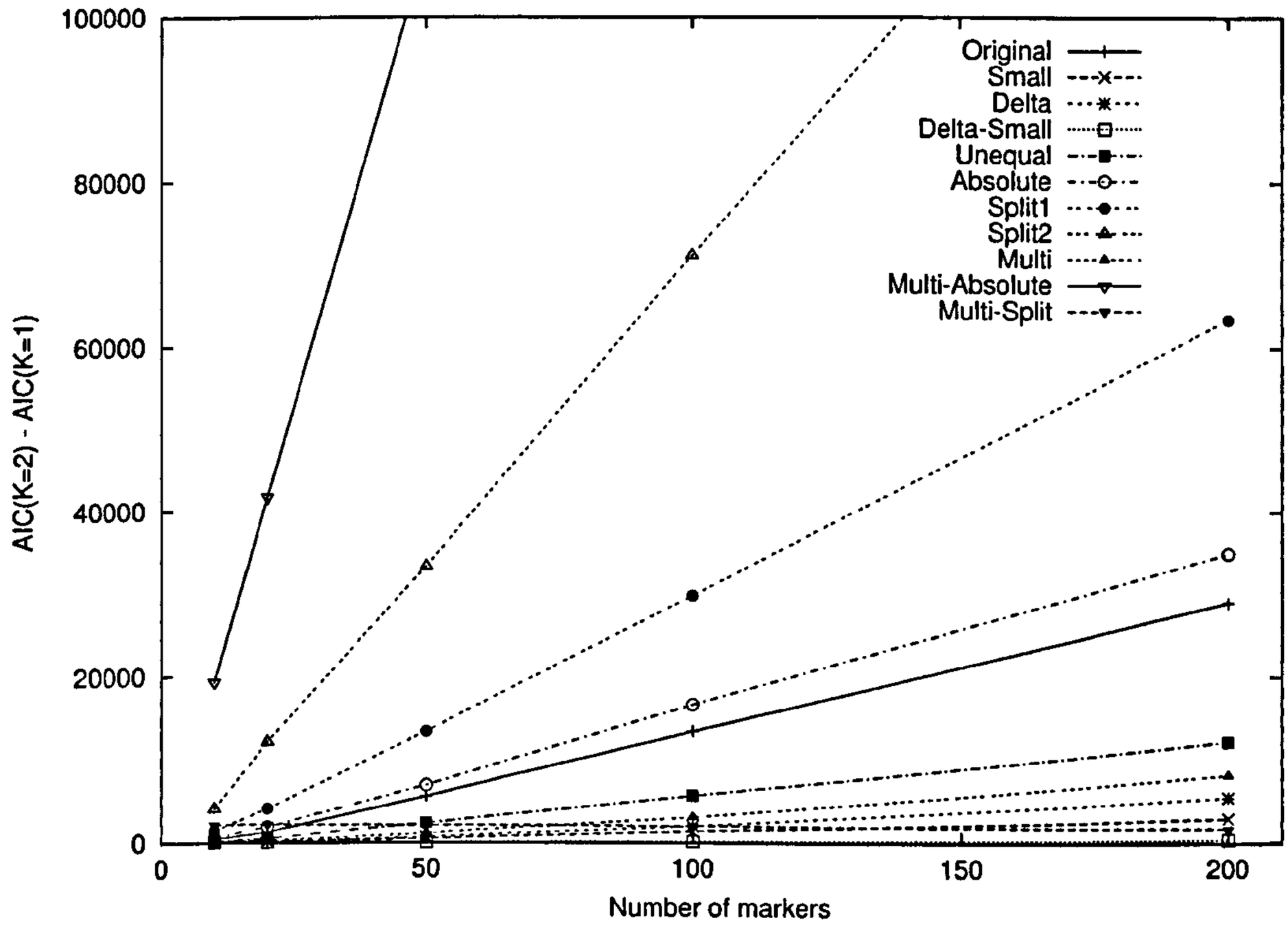


Figure 6.3: Simulations result: Δ_{AIC} for different models.

In all cases, the AIC difference appears to increase with increasing number of markers in a roughly linear manner. The 'Split' conditions resulted in increased ability to detect the two-class solution; using multi-allelic markers with allele frequencies as in the 'Multi-Absolute' condition had the greatest impact (note: the line goes off the scale). As expected, decreasing the number of markers, genetic distance between groups and sample size all result in reduced AIC differences.

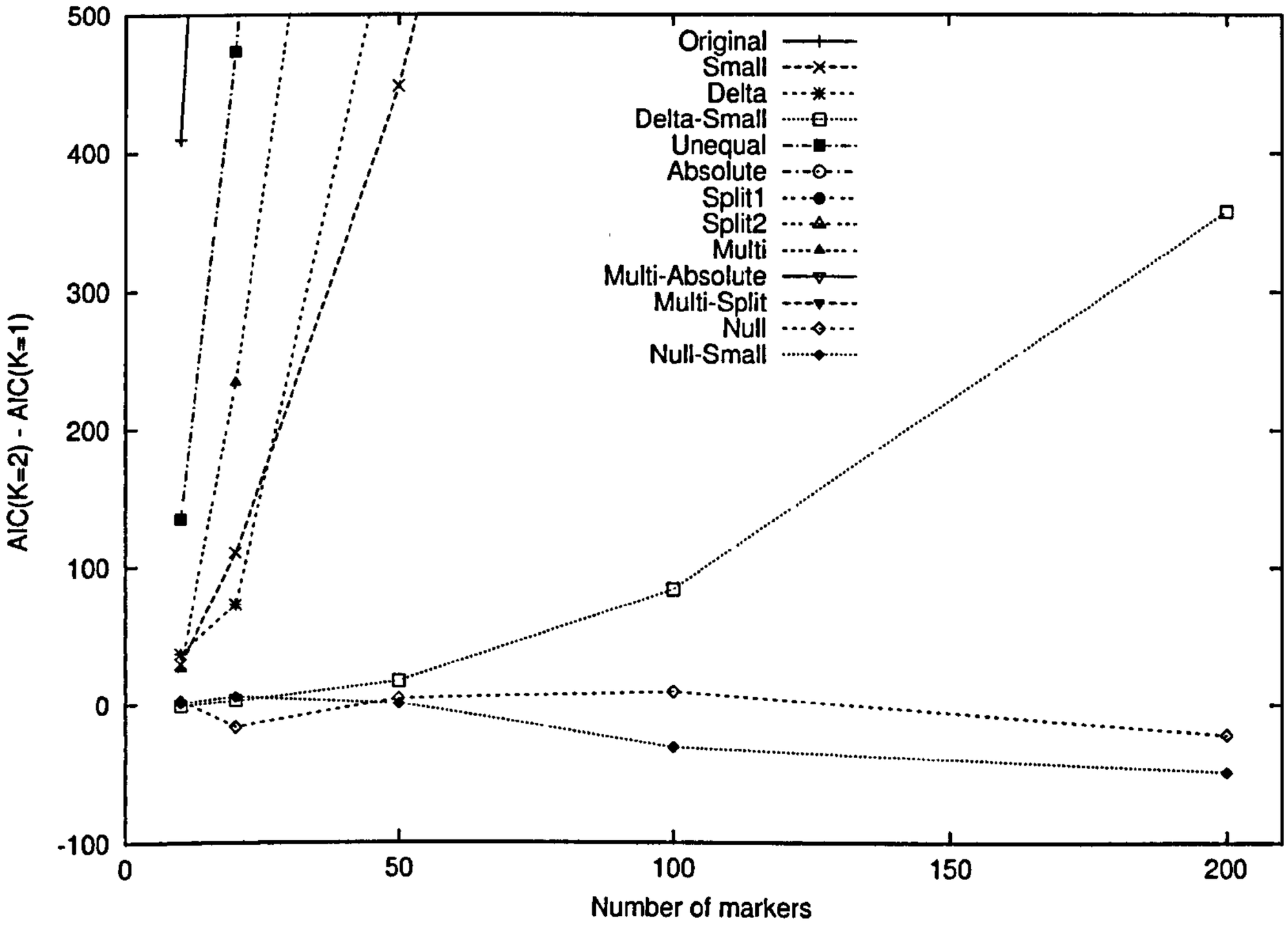


Figure 6.4: Simulation results: Δ_{AIC} for different models, reduced scale.

Figure 6.4 plots the same information, but changes the scale of the y-axis as appropriate for the conditions with little or no AIC difference. Although there is a trend for the AIC difference to become negative under the 'Null' conditions (and therefore represent a $K = 1$ solution) this is not as striking as the performance under the alternative. This plot also shows the poor performance of the 'Delta-Small' condition with fewer than 100 markers.

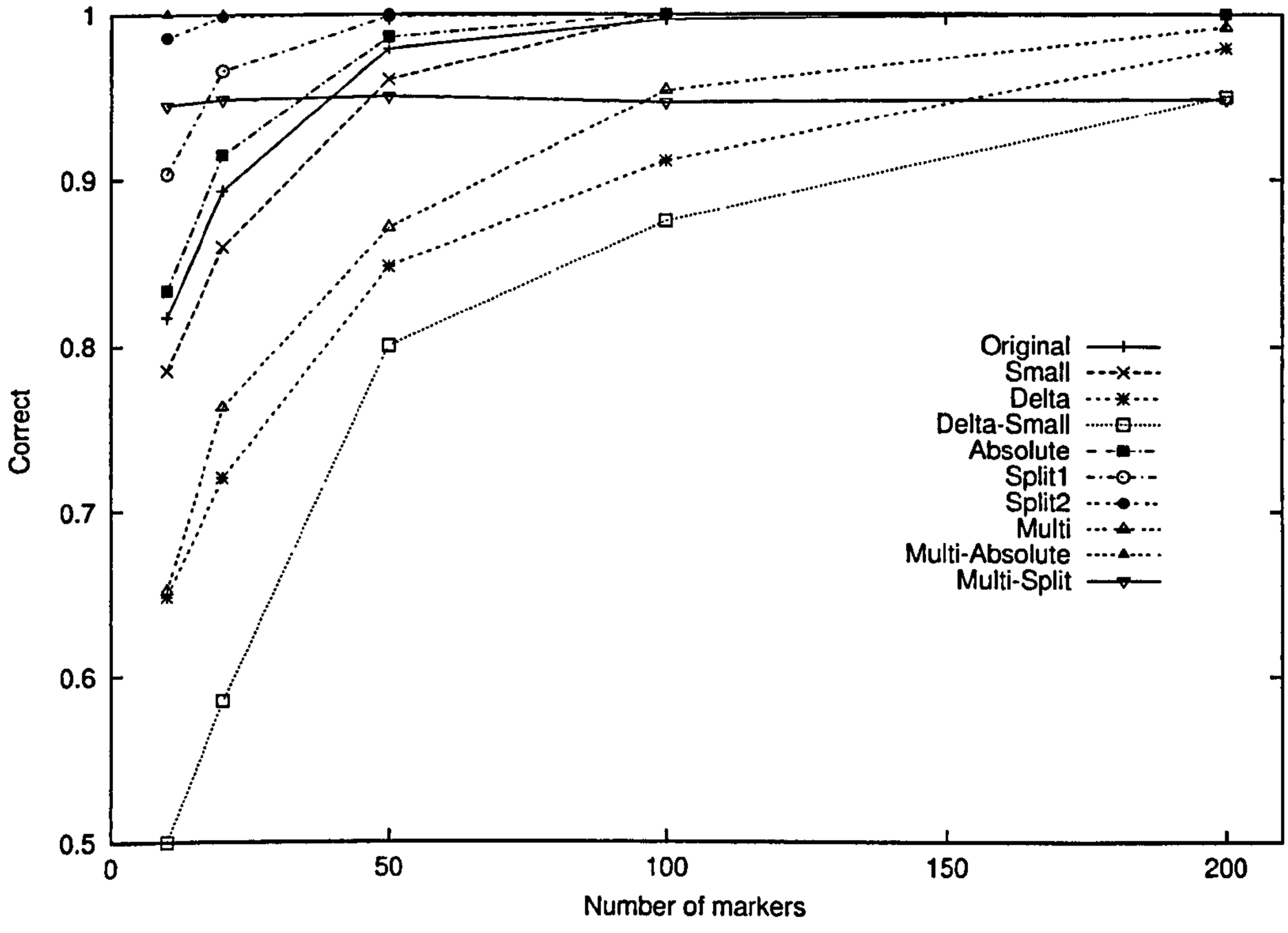


Figure 6.5: Simulation results: *Correct* for different models.

Finally, Figure 6.5 plots the classification accuracy rate for the different conditions by increasing M. Note that the line for the 'Multi-Split' condition is flat, as expected, as the number of informative markers does not increase with M. In most conditions, performance is acceptable with around 50 markers (above 95% accuracy) and near-perfect with around 200 markers.

6.4 Further simulations

6.4.1 Many subpopulations

All the previous datasets were simulated under a two-class (or as a homogeneous) model. In this section, samples with five different subpopulations are simulated. In the first instance, 1000 individuals (200 from each subpopulation) were generated

with 50 SNP markers. The marker allele frequencies were taken from 5 sets (i.e. 10 markers per set). The allele frequencies for the "1" allele in the five classes are listed for the five sets of markers, *A*–*E*:

Set	$P(G\|C=1)$	$P(G\|C=2)$	$P(G\|C=3)$	$P(G\|C=4)$	$P(G\|C=5)$
A	0.6	0.4	0.4	0.4	0.6
B	0.4	0.6	0.4	0.6	0.4
C	0.4	0.4	0.6	0.4	0.4
D	0.6	0.6	0.4	0.4	0.4
E	0.4	0.4	0.4	0.6	0.6

The sample was analysed for solutions $K = 1$ to $K = 7$: on the basis of lowest AIC, the $K = 7$ solution was the best-fit:

$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
104975.92	104516.51	104176.67	103861.84	103787.81	103779.09	103773.09

As can be seen, there is very little different in AIC between the $K = 5$, $K = 6$ and $K = 7$ models. In fact, plotting the AIC by K reveals a 'scree-plot' pattern, which gives support for the $K = 5$ solution (i.e. after this point the AIC 'levels off'). Two independent simulations are shown: although the mean AIC level is different, both show similar profiles.

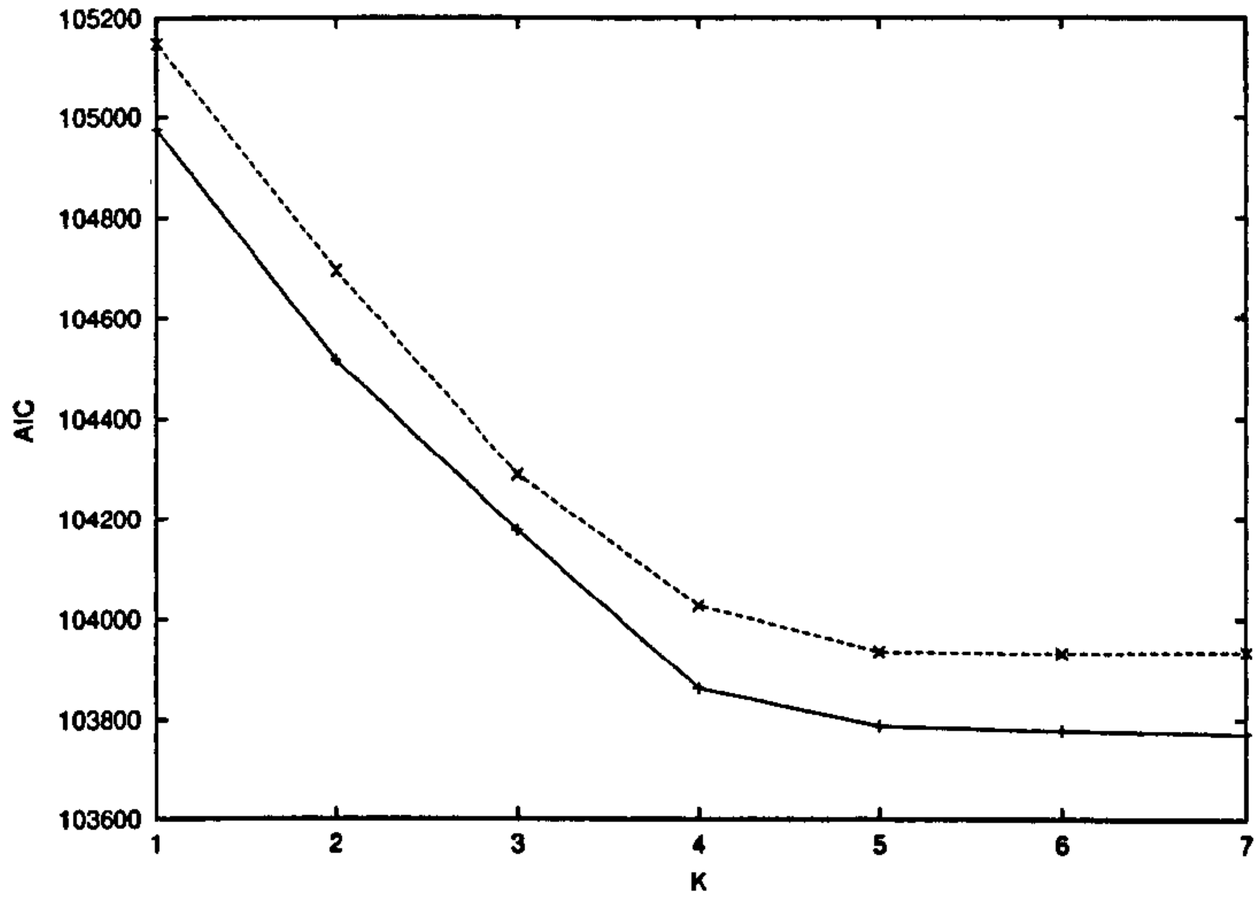


Figure 6.6: Simulations results: AIC plot by K for data with 5 subpopulations (for two independent datasets).

Examining the contingency table between the $K = 5$ solution and true subpopulation membership, there is a clear association between the estimated and true structure, although classification is far from a perfect:

	Estimated				
True	1	2	3	4	5
1	20	**154**	11	3	12
2	**153**	14	8	21	4
3	9	9	**159**	7	16
4	12	1	7	**140**	40
5	1	23	8	16	**152**

The $K = 6$ and $K = 7$ solutions appear to have picked off a small number of outlying individuals to form new classes: tabulating the prior class probabilities for the seven solutions shows that the $K = 6$ and $K = 7$ solutions look quite similar to the $K = 5$ solution:

K	$P(C)$	$\rightarrow$					
1	1.0000						
2	0.4034	0.5966					
3	0.2173	0.4769	0.3057				
4	0.3074	0.3023	0.2111	0.1791			
5	0.2007	0.1946	0.1929	0.1835	0.2283		
6	0.0263	0.1796	0.2222	0.1894	0.1959	0.1866	
7	0.0278	0.0605	0.2262	0.1854	0.1112	0.2007	0.1881

Repeating with 100 loci instead of 50, produces a similar result: K is still overestimated (a 6-class solution is favoured in this case). The characteristic scree-plot feature is still present (Figure 6.7 – again, two independent simulations show similar profiles). More investigation is needed to determine whether inspection of a scree-plot could be used to more accurately determine which solution to use.

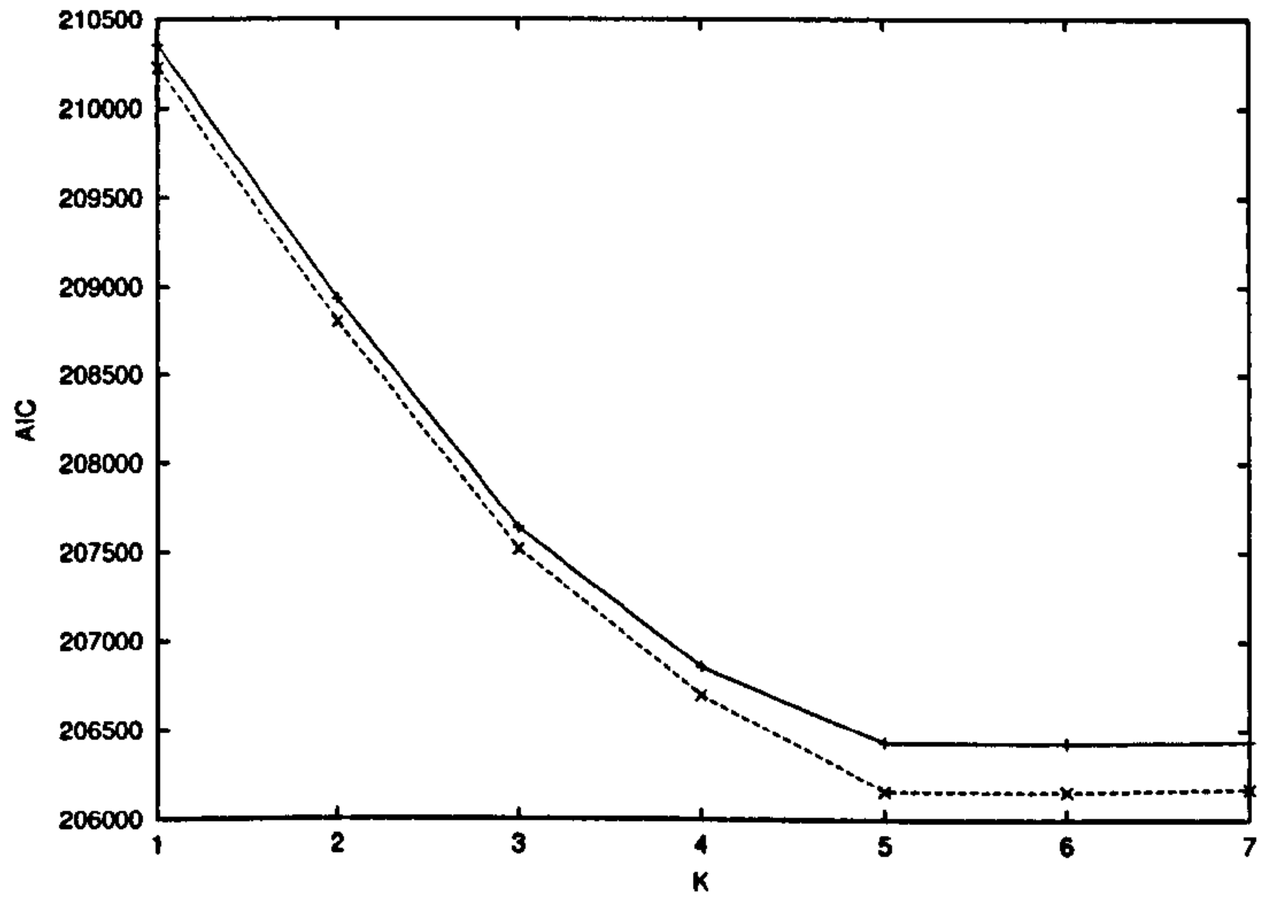


Figure 6.7: Simulation results: AIC plot by K for data with 5 subpopulations, 100 loci.

Classification performance under the 5-class solution with 100 loci has improved a little, as the contingency table shows:

	Estimated				
True	1	2	3	4	5
1	0	**194**	1	3	2
2	17	27	**154**	0	2
3	2	4	2	3	**189**
4	**184**	0	6	10	0
5	2	10	1	**186**	1

Of course, these simulations have not explored the different effects of marker type, genetic distance between group, etc, which where illustrated above. Whilst there is no reason to suspect that the impact of these factors would be any different when dealing with, say, 5 instead of 2 subpopulations, it does highlight the fact that the conditions chosen for the 5-subpopulation simulations are somewhat arbitrary – whether or not the conditions reflect what we might expect to find in practice is uncertain. The real-data applications shown below, however, do suggest that this approach can work well detect multiple subpopulations in a sample, as long as sufficient markers are typed.

6.4.2 Admixed subpopulations

This section illustrates the application of an admixed class model. Ninety SNP markers were simulated in 3 sets of 30 with the following ancestral-class-specific allele frequencies:

Set	$P(G\|C_A = 1)$	$P(G\|C_A = 2)$	$P(G\|C_A = 3)$
A	0.6	0.4	0.4
B	0.4	0.6	0.4
C	0.4	0.4	0.6

From the 3 ancestral classes, 5 derived classes were generated, each with 200 individuals. The matrix of admixture proportions was

$$\Theta = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \\ 0.50 & 0.00 & 0.50 \\ 0.50 & 0.50 & 0.00 \end{bmatrix}$$

which specifies 3 pure classes and two admixed classes. Both derived classes were simple 50:50 admixtures of two ancestral classes.

The AIC is tabulated below for a number of different solutions: no-admixture solutions $K = 1$ to $K = 4$ and 3 admixed solutions. Solution $K = 2 + 1$ represents 2 ancestral classes, 2 pure derived classes, one 50:50 admixed derived class. Solution $K = 2+3$ represents 2 ancestral classes, 2 pure derived classes and 3 admixed derived classes (25:75, 50:50 and 75:25 admixtures). Solution $K = 3+3$ represents 3 ancestral classes and 3 derived classes (i.e. the 3 50:50 admixture pairs between the 3 ancestral classes). The AIC values were:

$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 2+1$	$K = 2+3$	$K = 3+3$
188218.57	187112.51	186250.69	186238.38	187099.59	187101.96	186197.03

As expected, the $K = 3+3$ solution provided the best fit. The $P(C)$ for this solution are (with the admixture proportions also shown):

C_D	$P(C_D)$	Admixture proportions		
1	0.1920	1.0	0.0	0.0
2	0.1744	0.5	0.5	0.0
3	0.2229	0.0	1.0	0.0
4	0.0086	0.5	0.0	0.5
5	0.2096	0.0	0.5	0.5
6	0.1926	0.0	0.0	1.0

Note that only two of the admixed classes have sizeable $P(C)$ – derived class 4 is essentially empty and unnecessary, as expected. Although L-POP provides an automatic function to generate all possible admixture pairs, it is also possible to manually specify the Θ matrix. If the analysis is repeated, dropping the derived class 4, then the AIC is lower still (186195.21) because there are fewer parameters in this model.

Inspecting the contingency table of the estimated best-fit solution and true subpopulation membership reveals reasonable classification performance. Note that estimated classes "2" and "5" represent the admixed classes (i.e. corresponding to true classes "4" and "5"). Estimated class "4" is dropped.

	Estimated				
True	1	2	3	5	6
1	0	9	**172**	19	0
2	**172**	27	1	0	0
3	5	1	1	31	**162**
4	1	18	21	**133**	27
5	22	**113**	35	28	2

Figure 6.8 shows the multidimensional scaling plot from the best-fit solution (with class 4 dropped). The admixed classes have been shaded in gray: note how these are halfway between the pure classes they are an admixture of. The three pure classes are equidistant, forming a triangle. If derived class 4 had of featured in the solution,

it would appear halfway between classes "1" and "6". (Remember that the estimated class labels are arbitrary, e.g. estimated class "5" corresponds to true class "4", which is an admixture of true classes "1" and "3", or estimated classes "3" and "6". Of course, these complications will not arise when analysing real data.)

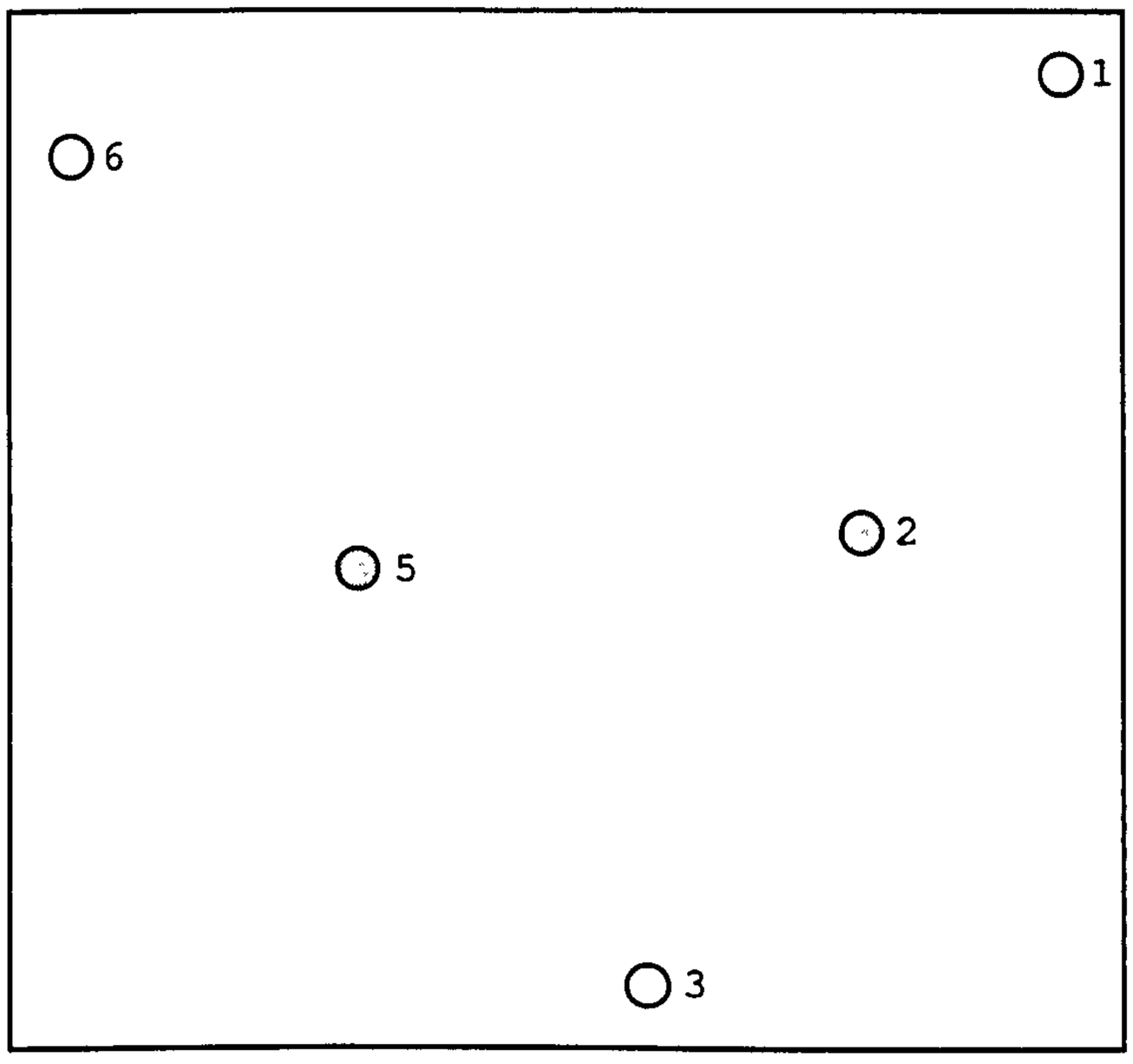


Figure 6.8: Simulation results: Multidimensional scaling plot for simulated dataset with admixture.

More work is needed to determine the utility of the admixture approach. One difference between STRUCTURE and L-POP is that the former allows continuously varying admixture proportions to be estimated for each individual. This is not possible within a LCA framework (i.e. there must be a finite number of classes, which would not be the case if every individual could have unique admixture proportions). In reality, allowing for a resolution of up to 25% admixture is probably satisfactory. This resolves ancestry down to the grandparental level (i.e 25:75, or 25:25:50). One approach would be to run a model with all possible admixture combinations at the 25% level, and then drop the admixed classes that have $P(C)$ below some threshold (i.e. as for

the example above, where derived class 4 was dropped). The analysis of the Wilson *et al* data reported below illustrates an real-life example of admixed subpopulations.

6.4.3 The Hardy-Weinberg equilibrium assumption

Population stratification is not the only cause of Hardy-Weinberg disequilibrium, as mentioned above. One other potentially common cause is selective genotyping error. Consider the scenario in which heterozygous individuals are more likely to have a missing genotype, as can in practice happen. This loss of heterozygosity is not likely to lead to spurious association – but might it lead to "spurious stratification". That is, the current method might take HWD due to missing heterozygous genotypes as evidence of stratification and therefore favour a spurious $K > 1$ solution.

This possibility was investigated by simulation: 10 replicate homogeneous datasets of 400 individuals and 40 SNPs (equal allele frequencies) were simulated. Heterozygotes were designated missing at probability 0%, 25%, 50% and 75%. Therefore, in the last (unrealistically extreme) condition (75% missing) substantial deviations from HWE were observed.

The data were analysed for $K = 1$ and $K = 2$ solutions in the standard manner. The approach was also modified to allow `L-POP` to relax the within-class HWE assumption, by treating genotypes as the unit of response rather than alleles (i.e. equivalent to assuming all individuals to be haploid and that each genotype is a unique allele).

	$AIC(K=1) - AIC(K=2)$	
Missing A_1A_2	HWE assumed	HWE relaxed
0%	-72.63	-68.01
25%	-52.25	-60.78
50%	13.41	-69.79
75%	119.76	-72.55

Table 6.16: The impact of selective genotyping failure: relaxing the within-class HWE assumption.

As Table 6.16 illustrates, a large percentage of the heterozygotes must be missing in order to favour a two-class solution (i.e. positive values of $AIC(K=1)-AIC(K=2)$) – it is very unlikely that this level of genotyping failure would occur in practice for all markers. Also, the specification of equal allele frequencies represents a 'worst-case scenario' (i.e it gives the highest possible frequency of heterozygotes). Furthermore, when the option to relax the within-class HWE assumption was implemented, the AIC difference remained invariant to the marker HWD.

In summary, it is unlikely that failure to genotype heterozygotes could generate spurious evidence for stratification. In any case, the method can relax this assumption. Most of the signal for stratification comes from LD rather than HWD, and so, most probably, little information will be lost even in the presence of true stratification when relaxing the HWE assumption.

6.5 Data applications

6.5.1 Satten *et al* data

Satten et al. (2001) applied their LCA method to a simulated dataset based on allele frequencies for 12 multi-allelic short tandem repeat (STR) loci in Argentinian and

Native American samples, successfully recovering the simulated class structure. This section reports the attempt to replicate this results using L-POP.

Table 6.17 gives the allele frequencies used to simulate the sample of 250 individuals. The four populations were simulated in the ratio 1:1:1:7. Running L-POP for $K = 1$ to $K = 5$, the best-fit solution on the basis of AIC was for $K = 4$ (the AIC values for the five solutions were 14151.288, 12702.709, 12205.923, 11971.869 and 12023.728 for $K = 1$ to $K = 5$ respectively). In fact, this solution perfectly recovered the simulated population substructure: the four prior class probabilities were exactly 0.1, 0.1, 0.1 and 0.7 and all posterior probabilities were either 0 or 1 (to four decimal places), with the individuals correctly grouped according to simulated population. Table 6.17 also shows the locus-specific genetic distances (in parentheses under the locus name), which is an index of the informativeness of that marker in the final solution. These values are mostly quite high reflecting the substantial degree of among-class variation in allele frequency. This seems a particularly easy problem mirroring the previous 'Multi-Absolute' simulation. These results suggest the utility of using highly polymorphic loci, which are likely to have allele frequencies of 0 in some subpopulations.

6.5.2 Pritchard *et al* data

Simulated data used in Pritchard and Donnelly (2001) to assess the performance of structured association as compared to genomic control was obtained from the author and analysed using L-POP. Although Pritchard and Donnelly (2001) also looked at the behaviour of the correction for stratification in a case-control test of association, this section limits the analyses to the detection of stratification only.

Pritchard and Donnelly (2001) drew individuals from three distinct populations, at fixed frequencies 116, 117 and 167. The subpopulation allele frequencies were modelled using the Balding and Nichols (1995) model, in which the allele frequency at locus l in

Locus (Info)	Population European	Mapuche	Tehuelche	Wichi
FABP	0.589	0.683	0.732	0.485
(0.408)	0.110	0.058	0.107	0.162
	0.300	0.260	0.161	0.353
CSF1P0	0.33	0.266	0.339	0.226
(0.383)	0.313	0.282	0.232	0.194
	0.298	0.367	0.411	0.581
	0.059	0.085	0.018	0.000
D6S366	0.082	0.091	0.143	0.000
(0.597)	0.204	0.114	0.071	0.000
	0.277	0.341	0.446	0.557
	0.119	0.136	0.036	0.086
	0.091	0.125	0.036	0.029
	0.183	0.159	0.143	0.200
	0.028	0.011	0.018	0.071
	0.015	0.023	0.107	0.057
F13A	0.151	0.222	0.357	0.173
(0.421)	0.060	0.122	0.125	0.077
	0.202	0.122	0.054	0.346
	0.209	0.178	0.143	0.115
	0.325	0.344	0.304	0.288
	0.053	0.111	0.017	0.000
FES	0.260	0.170	0.143	0.257
(0.486)	0.420	0.500	0.714	0.543
	0.247	0.284	0.107	0.043
	0.073	0.045	0.036	0.157
TH01	0.233	0.526	0.286	0.132
(0.610)	0.250	0.298	0.429	0.721
	0.105	0.088	0.018	0.000
	0.185	0.026	0.089	0.015
	0.226	0.140	0.179	0.132
HPRTB	0.032	0.000	0.000	0.000
(0.722)	0.179	0.032	0.091	0.000
	0.317	0.323	0.227	0.357
	0.285	0.403	0.591	0.167
	0.137	0.242	0.091	0.357
	0.050	0.000	0.000	0.119
vWA	0.063	0.096	0.036	0.014
(0.593)	0.099	0.077	0.054	0.014
	0.294	0.577	0.429	0.514
	0.297	0.125	0.214	0.343
	0.246	0.212	0.268	0.114
D13S317	0.090	0.020	0.000	0.000
(0.560)	0.160	0.240	0.150	0.464
	0.060	0.070	0.050	0.179
	0.290	0.120	0.150	0.089
	0.250	0.260	0.300	0.089
	0.100	0.180	0.225	0.179
	0.040	0.110	0.125	0.000
D7S820	0.156	0.070	0.050	0.000
(0.540)	0.115	0.050	0.050	0.070
	0.276	0.220	0.175	0.125
	0.245	0.420	0.525	0.450
	0.159	0.210	0.200	0.250
	0.046	0.030	0.000	0.105
D16S539	0.156	0.110	0.225	0.125
(0.304)	0.100	0.130	0.075	0.232
	0.294	0.240	0.100	0.321
	0.159	0.370	0.550	0.250
	0.195	0.150	0.050	0.071
RENA-4	0.772	0.719	0.881	0.690
(0.613)	0.074	0.229	0.023	0.000
	0.153	0.041	0.095	0.310

Table 6.17: Allele frequencies for 12 STR markers from Argentinian and Native American populations (from Satten et al. (2001)).

subpopulation i was simulated from a beta distribution with parameters $\{(F_i p_i/(1-F_i)), F_i(1-p_i)/(1-F_i)\}$, where $i \in \{1,2,3\}$. The ancestral allele frequency was drawn from a uniform distribution in (0.1,0.9). Wright's F_{ST} for the three subpopulations was set to 0.01, 0.02 and 0.04 respectively.

Pritchard and Donnelly (2001) simulated either 50, 200 or 1000 diallelic loci for the 400 individuals. The condition with 1000 loci was omitted from the present study for two reasons. First, as is shown below, 200 loci provided ample information to detect all three subpopulations, so 1000 loci would, in this case, be an overkill. Second, L-POP ran into computational difficulties when dealing with 1000 loci. Although it should be easy to extend the limits in future versions of L-POP, if one were to have a thousand or more loci at hand, a better strategy might be to split the loci up into two or more datasets, in order to provide an internal validation of the solution.

The parameters as set correspond to a λ inflation factor of 1.24 – a modest amount of stratification. Simulating 20 replicate samples under each condition, with 50 loci STRUCTURE was unable to detect the presence of three distinct subpopulations: for 4 datasets $K = 1$, for the other 16 $K = 2$. With 200 (or 1000) loci, a three-class solution was correctly selected every time, however. The results using L-POP differ somewhat. Although classification with 200 loci is near perfect, as with STRUCTURE, the behaviour of the model with only 50 loci seems to differ, as Table 6.18 shows. Whilst STRUCTURE under-estimates K, L-POP correctly estimates K six times, but otherwise tends to over-estimate K at 4. (Note: the data obtained from Pritchard and Donnelly (2001) only contained 19 replicates of the 50 loci condition.)

Comparing the classification of individuals according to the L-POP solutions against true subpopulation membership for the $M = 50$ simulations, there was a highly significant relationship in every case (the average χ^2 values were 228.3, 200.6 and 191.1 for the two, three and four class solutions respectively). Using a more appropriate metric, however, the adjusted RAND coefficient between the true solution and the estimated

		K			
M	Method	1	2	3	4
50	STRUCTURE	4	16	0	0
	L-POP	0	1	6	12
200	STRUCTURE	0	0	20	0
	L-POP	0	0	20	0

Table 6.18: Comparison of STRUCTURE and L-POP solutions. Note: only 19 replicates for the L-POP condition with 50 loci.

solutions, the average value was 0.256, which is quite low. Although the solutions were clearly not independent of true subpopulation structure, the classification was typically not precise. For example, one of the 6 three-class solutions from the 50 loci condition (largest cell frequencies in **bold** type):

	Estimated		
True	1	2	3
1	10	**73**	33
2	6	**108**	3
3	**131**	30	6

In this case, the true subpopulation "3" with the highest F_{ST} (0.04) has been separated out from "1" and "2" – it corresponds to estimated class "1". True subpopulations "1" and "2" have both been pooled in estimated class "2" however, although some individuals from true subpopulation "1" have split off to form the third estimated class, "3". This classification is therefore a partial success, in that it has captured the main essence of the structure within the sample (true subpopulation "3" is the largest and most distant of the three subpopulations: therefore "1 & 2" versus "3" is the primary axis of stratification.)

An example of one of the 12 four-class solutions is:

	Estimated			
True	1	2	3	4
1	23	56	28	9
2	14	18	75	10
3	50	13	17	87

In this case, the solution has done a better job of discriminating between true subpopulations "1" and "2" but subpopulation "3" has been split into two separate estimated classes ("1" and "4").

With 200 loci performance was much better, with an average adjusted RAND coefficient was 0.84514 and average classification accuracy of 94% on the basis of highest posterior probability. To compare these results to the previous simulations, performance appears to be worse for the equivalent number of markers. The main reasons for this would appear to be (1) the higher number of subpopulations (2) the smaller degree of genetic separation between two of the subpopulations and (3) the smaller sample size in the Pritchard and Donnelly (2001) data.

6.5.3 Wilson *et al* data

Wilson et al. (2001) presented a stratification analysis of a sample collected from eight geographically diverse regions, using STRUCTURE, to study the population genetic structure of variable drug response. The sample consisted of 354 males from 8 ethnic labels: Afro-Caribbean, Bantu, Ethiopian, Norwegian, Armenian, Ashkenazi, Chinese and Papuan New Guinea. The sample was typed on 38 micro-satellite markers (16 autosomal and 22 on the X chromosome). STRUCTURE recovered a four-class solution, roughly corresponding to

A Norwegian, Armenian, Ashkenzai, Ethiopian

B Bantu, Afro-Carribbean, Ethiopian

C Chinese

D Papuan New Guinea

Wilson et al. (2001) noted that common ethnic labels such as "Black" or "Asian" would therefore fail to capture the true population structure as outlined in their analysis. The label "Black", for example, would fail to capture the distinction between sub-Saharan Africans and North Africans — the latter have largely descended from Caucasian ancestry (explaining why Ethiopians cluster with the Europeans as well as the sub-Saharan Africans). Likewise, the label "Asian" fails to discriminate between Chinese and Papuan New Guinea. However, these conclusions have questioned by Risch et al. (2002), in the context of comparing self-reported racial labels to genetically-defined clusters (discussed below).

The STRUCTURE solution and the L-POP solution for $K = 4$ are very similar, with an adjusted RAND coefficient of 0.82. (In contrast, the adjusted RAND coefficients between the four-class cluster solutions and the eight-category self-reported ethnicity are around 0.3.)

	STRUCTURE			
L-POP	1	4	3	2
4	**173**	0	1	2
1	5	**35**	1	6
2	3	3	2	**42**
3	6	1	**74**	0

The best solution generated by L-POP is one which allows for admixture, between the European and sub-Saharan African groups, to describe the Ethiopians:

L-POP	Ban	Ash	Eth	Nor	Arm	Chi	PNG	AfCa
1	**44**	0	2	0	0	0	0	**22**
2	0	0	0	0	1	**36**	0	0
3	0	0	0	0	0	0	**45**	0
4	0	**46**	**15**	**46**	**42**	1	0	3
1&4	2	0	**31**	0	1	0	0	4

It is important to remember that admixture only implies that a class has, on average, allele frequencies that are at intermediate levels between two or more other classes. If C is an admixed class deriving from ancestral classes A and B, this says nothing about the population genetic history of the three classes. That is, the data are equally consistent with a 'merging' event (A and B result in C) as with a 'splitting' event (C results in A and B).

Figure 6.9 shows the results of applying multi-dimensional scaling to the inter-class genetic distance matrix, for an eight class solution (left panel) as well as the best-fit solution (right panel). There is only a partial correspondence between self-reported ethnicity and class in the eight-class solution; the best-fit solution shows a clearer picture, with the Ethiopian admixed class half-way between the European and sub-Saharan Africans in this space.

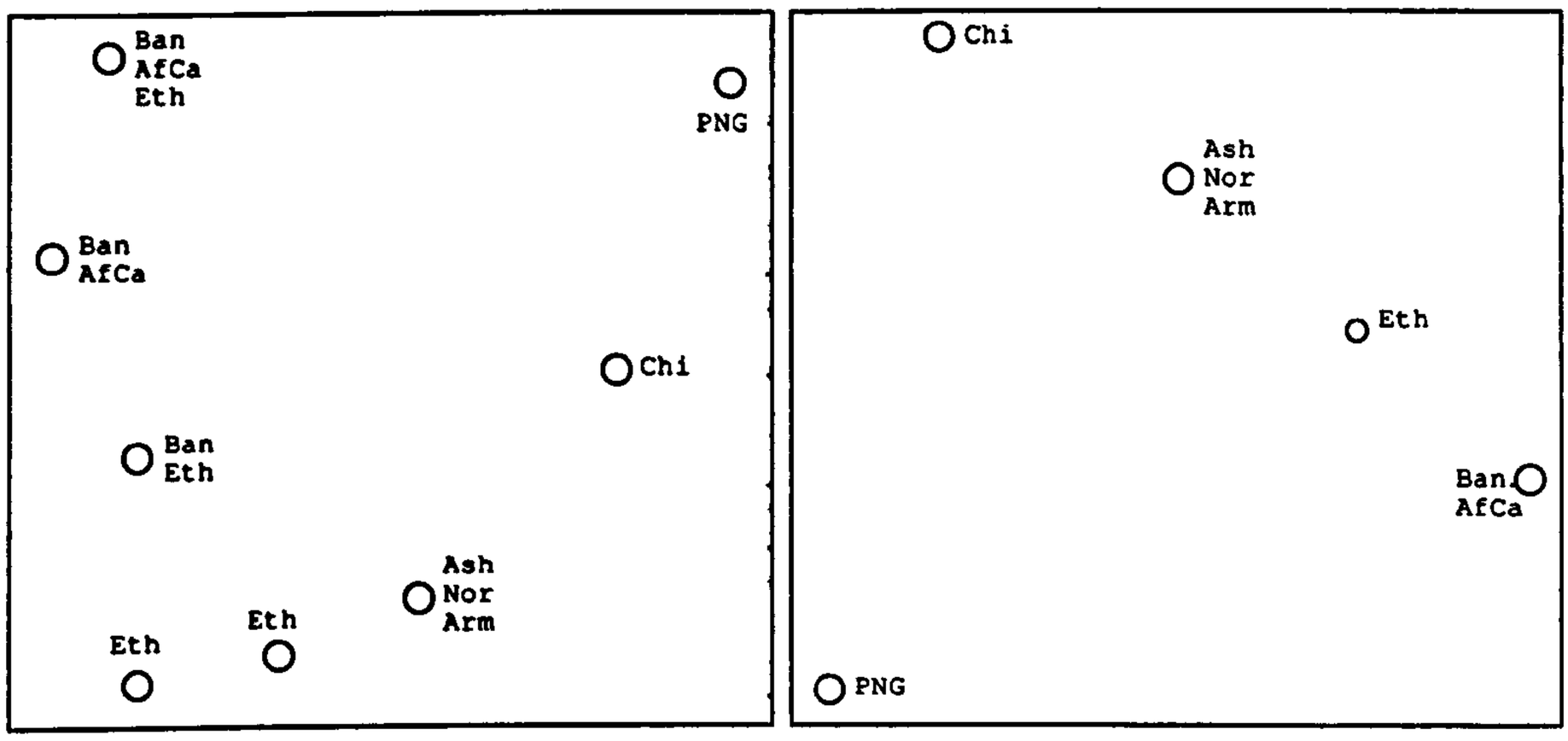


Figure 6.9: Multidimensional scaling plot for $K = 8$ (left) and $K = 4 + 1$ (right) solutions for Wilson et al. (2001) data.

Finally, examining the distribution of posterior probabilities $P(C|G)$ for the four-class solution reveals a difference between approaches: as shown in Figure 6.10, values for $P(C|G)$ are much closer to either 0 or 1 from L-POP than from STRUCTURE, based on the four-class solution. Whether this pattern is indicative of a consistent difference between the two approaches needs further investigation, however.

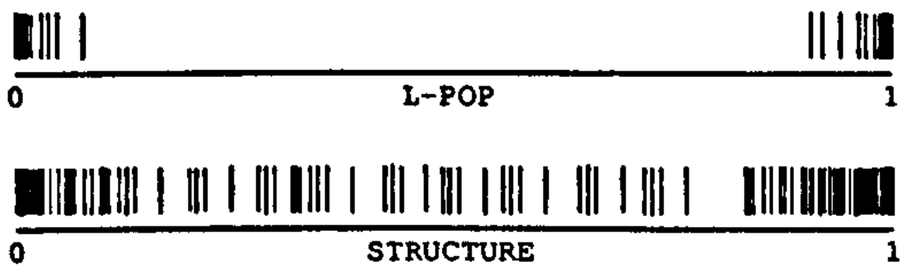


Figure 6.10: Distribution of $P(C|G)$ for L-POP and STRUCTURE.

6.5.4 Dunedin sample

A final application of L-POP used data from the Dunedin Multidisciplinary Health and Development Study (Silva and Stanton, 1996). This sample consisted of 953 unrelated individuals (84 individuals who had mostly missing genotype data, or who were part of a twin pair, were removed from the original sample). As well as genotypes

for 13 SNP markers and a microsatellite, the majority of individuals also had self-reported information on grandparental ancestry (European versus non-European). In particular, the non-European ancestry indexed Maori ancestry (Dunedin is in New Zealand).

Based on self-reported grandparental ancestry, approximately 10% of the sample is of (partial) non-European descent; around 5% of the alleles have descended from non-European individuals. Dummy variables were created, E, NE1 and NE2:

E	No non-European grandparents	856
NE1	At least one non-European grandparent	93
NE2	At least two non-European grandparents	50

NE1 is approximately 10% of the sample, NE2 is approximately 5% (i.e. NE2 is a subset of NE1). Calculating the proportion of individuals' genomes that are of European descent indicates the following distribution:

% European	N
0	12
0.25	2
0.5	36
0.75	43
1	856

As mentioned above, the initial dataset contained genotype information on 14 markers. Based on all the markers, a three class solution was extracted using L-POP; prior class probabilities were 21%, 51% and 28%. However, closer inspection (using the LOCINFO option to generate locus-specific genetic distances) revealed that only a subset of the loci were contributing to the solution: loci 1, 3, 12 and 14. Subsequently, it became clear that several of the markers were linked and that these linked pairs were generating a spurious solution: an important precondition is that markers are unlinked, as the signature of stratification is that unlinked markers show link-

age disequilibrium. Having linked markers showing linkage disequilibrium gives no information about stratification and, in this case, gave a spurious solution.

Markers 1 and 3 are linked; markers 12, 13 and 14 are all linked. This is illustrated in very high marker-marker LD between 1–3 and 12–14. (Marker 13 is not in LD with either 12 or 14.) Markers 1, 13 and 14 were excluded from all subsequent analysis, to remove the LD due to linkage. There is still significant LD between other pairs of more distant, unlinked markers, however, notably 2–7, 3–7 and 2–11.

For the 13 SNPs, Table 6.19 shows the allele frequencies, separately for each self-reported ancestry group. Inspecting these frequencies suggests that differences exist between the European and non-European groups. In particular, the frequencies for NE1 are typically in between E and NE2, as expected under stratification (marker 6 is a particularly clear example). Between E and NE2, for the 13 SNPs, there is an average absolute difference in allele frequency (i.e. δ value) of 0.1. Based on the previous simulations, one might not expect great power to detect this stratification: the sample size is moderate (≈ 1000) but the δ value is not very large, the subpopulations are unequal in size, and there are only 11 SNP markers.

Marker	E	NE1	NE2
1*	0.75	0.85	0.86
2	0.79	0.87	0.88
3	0.70	0.76	0.82
4	0.69	0.67	0.69
5	0.65	0.73	0.73
6	0.29	0.38	0.46
7	0.87	0.73	0.68
8	0.41	0.35	0.31
9	0.58	0.63	0.65
10	0.68	0.77	0.81
11	0.60	0.64	0.69
12	0.66	0.64	0.59
13*	0.77	0.83	0.88

Table 6.19: Frequency of '1' coded allele for the 13 SNPs by ancestry group. The two SNPs marked * were excluded from the analysis, along with the micro-satellite marker (not shown here).

For the revised dataset containing the 11 unlinked SNPs, evidence for population stratification was detected, with a $K = 3$ solution giving the lowest AIC value. Table 6.20 shows the AIC for a number of different models, including two admixture models. The 2 + 1 model represents 2 pure ancestral classes and a single admixed class that is a 50:50 mixture of the ancestral classes. The 2+3 model represents 2 pure classes and three admixed classes: 25:75, 50:50 and 75:25 admixtures. Each model was repeated 100 times with different random starting values, to ensure that the final solution represents a global and not local minimum.

K	AIC
1	17697.8167
2	17686.6367
3	17678.1928
4	17683.9790
5	17693.1932
2+1	17684.1893
2+3	17688.2685

Table 6.20: AIC values for different K in the Dunedin sample.

The differences in AIC between the different solutions are all quite modest, representing the low resolving power of the analysis. In this situation, one would not necessarily place a lot of confidence on the $K = 3$ solution and its corresponding classification of individuals, based on the performance in the simulation studies. The prior class probabilities for the $K = 3$ solution were 0.0511, 0.0535 and 0.8953. Based on highest posterior probability criterion, the assignment to classes did not match self-reported ancestry particularly closely:

	Estimated		
Self-Report	1	2	3
0	17	25	814
1	2	1	40
2	6	3	27
3	1	0	1
4	6	0	6
.	0	1	3

Neither did the $K = 2$ or $K = 4$ solutions offer a more straightforward interpretation of the data. If $\delta \approx 0.1$ is representative of the genetic distance between groups of European and Maori ancestry, then more markers would be required in order to reliably detect these two groups. Alternatively, focusing on just a few markers known to have markedly different allele frequencies between groups would help.

6.6 Summary

6.6.1 Power issues

The new genetic background methods (genomic control and structured association) still require a fully comprehensive evaluation of power issues. This has been partially attempted, for example, by the Pritchard and Donnelly (2001) simulations reported above and Bacanu et al. (2000) on the power of genomic control versus the TDT. Bacanu et al. (2000) found that in the absence of stratification, genomic control approaches are more powerful, especially with common diseases. However, in the presence of stratification the results are more complex, although genomic control methods seem preferable when the level of stratification is quite low. Overall, it seems that these methods can work with as few as 20 loci – this figure seems roughly supported by the simulations presented in this chapter.

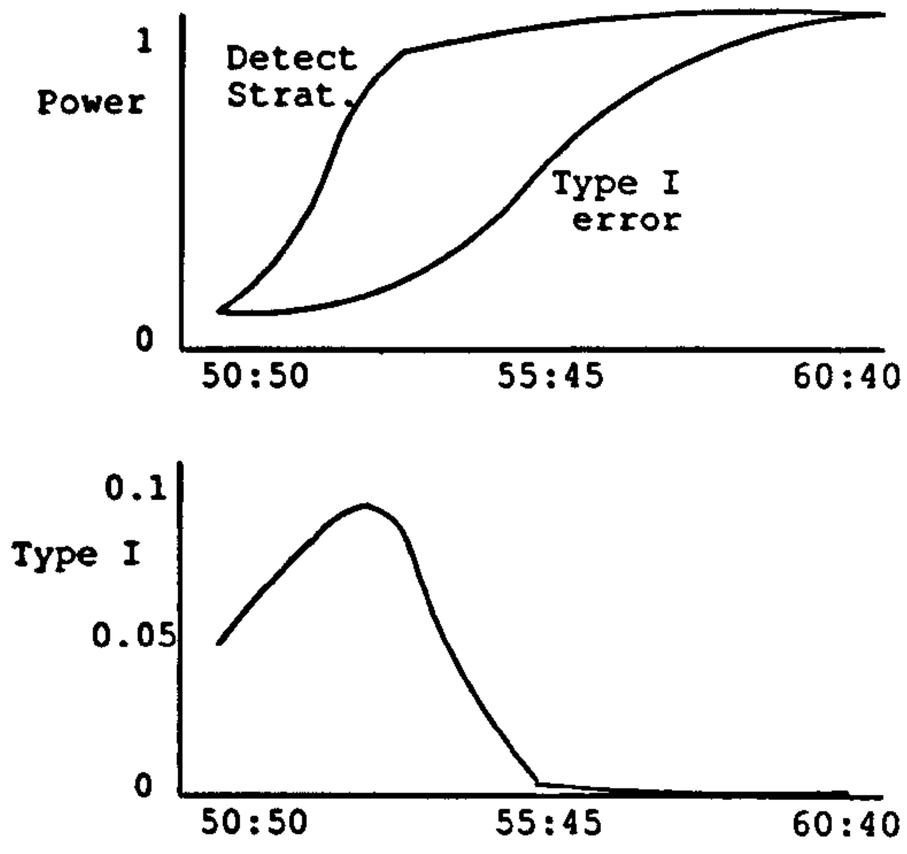


Figure 6.11: Power issues in GC (after Cardon and Bell (2001)).

Unlike family-based methods, which logically control for stratification, genetic background methods only probabilistically control for stratification. That is, whether or not the stratification is correctly detected in the first place is subject to certain power constraints. Cardon and Bell (2001) point out some potential problems that can result from less-than-perfect ability to detect low levels of stratification. Figure 6.11, in the top panel, shows the relative power of genomic control to detect stratification versus the type I error resulting from spurious association when population stratification is not controlled for. The x-axis represents increasing magnitude of stratification effect. The Figure appears to suggest that all is well, as power to increase stratification rises more quickly with increasing stratification than the impact on type I error rate. However, as the bottom panel shows, the overall type I error rate can still be inflated (doubled) at low levels of stratification (i.e. when the power of the genomic control method is significantly less than 100%). These results suggest that genetic background methods do not provide absolute protection against stratification.

However, one great advantage of the structured association approach is that cluster-membership can become a variable in analysis to do more than just control for stratification effects. For example, it is possible to look for cluster-based interaction effects

that might represent $G \times E$ (i.e. E is indexed by cluster) or epistasis (i.e. often called a 'genetic background' effect) or allelic heterogeneity (different alleles promote risk in different clusters).

6.6.2 Self-reported race versus genetically-defined clusters

The utility of genetically-defined clusters as opposed to self-reported ethnicity has recently been called in question by Risch et al. (2002). The article notes that the vast majority of human population genetic studies have identified substantial genetic differences between race, and that these differences cluster are strongest when defined on a continental basis, giving five major clusters: Africans, Caucasian, Pacific Islander, East Asian and Native American. Although the continental boundaries must be modified slightly to account for migrational patterns that have blurred these geographical boundaries, the continental organisation is, from an evolutionary perspective, unsurprising.

Risch et al. (2002) note that the four-class solution of Wilson et al. (2001) adheres to the continental scheme (they also note that most population geneticists would not classify Chinese and Papuan New Guinea together as "Asian" but would use the continental definitions of "East Asian" and "Pacific Islander" instead). Risch et al. (2002) point out that the Wilson et al. (2001) analysis fails to distinguish between Norwegians, Ashkenazi Jews and Armenians despite numerous studies showing genetic differences between these groups, whereas self-reported ancestry presumably would have. Another example of self-reported ethnicity being more accurate than genetically-defined grouping involves the Pima Indian / Caucasian admixture result: self-reported admixture correlated more strongly with diabetes than clusters defined genetically using 18 standard blood markers (Williams et al., 2000).

It could be that more loci are needed in order to more accurately differentiate between races on an intra-continental level, or to differentiate between different degrees

of inter-continental admixture. In contrast, to separate two populations with ancient separation and no migration would require far fewer loci. Risch et al. (2002) present a simple method of quantifying the number of diallelic loci needed, based mainly on the average δ value between two populations.

Assuming equal representation of the two populations, and an average δ of 0.2, Risch et al. (2002) calculate that 115 markers are needed for a misclassification rate of $1/1,000$; 218 markers are required for a misclassification rate of $1/100,000$. For an average δ of 0.1, the number of markers required are 474 and 901, respectively. These results seem a little high, compared to the simulations conducted here; also, demanding a misclassification rate of $1/100,0000$ seems a little excessive.

Recent large-scale studies have observed the distribution of δ between various ethnic groups (Dean et al., 1994; Smith et al., 2001). For SNPs, these typically have median values around 0.2 between the major ethnic groups; the top 20% have median values between 0.4 and 0.5. The corresponding values for multi-allelic markers are typically slightly higher. These values would suggest that genetic background approaches will easily discriminate between major ethnic groups with a reasonable number of markers (e.g. 50) and acceptable misclassification rates (a rate of 1/1000 seems completely reasonable).

What the distribution of δ is likely to be between different within the same major ethnic race is less clear. However, as the present simulations have shown, there are a number of other factors that influence power to detect stratification above and beyond average δ value. Ultimately, whether or not self-reported ethnicity provides a better index is something that can and should be empirically validated – indeed, both should be used when available.

6.6.3 Future directions

There are several future directions for the development of this method. First, it will be important to solve certain computational issues inherent in the application of the maximum-likelihood E-M framework, specifically the problem of local minima and constraints on the number of marker loci practicable.

It might be desirable to extend this method for use with full-sibships. This would involve estimating the parental genotypes based on the sibship genotypes and population allele frequencies (all members of a full sibship belong to the same class, by definition). Combined with the between-within sibship association model (Fulker et al., 1999), an index of subpopulation membership could be used as a covariate to increase the power of the between-sibships component (the within-sibship component is already robust to population stratification effects).

Admixed classes currently represent what might be called 'ancient' admixture: the model assumes that two classes intermingled many generations ago, and have since settled down such the HWE is restored. Another model of admixture is to assume that an individual is the first generation offspring of parents from different ancestral classes. This would involve modifying the M-step such that the product terms are of the form $P(G|C = Paternal)P(G|C = Maternal)$ and also changing the allele counting in the E-step.

A very important issue is the selection of an optimal marker set – several studies have begun to look at divergence in allele frequency for many markers between the major ethnic groups (Cargil et al., 1999; Halushka et al., 1999). Using these markers would be preferable for two reasons: first, they provide the greatest discriminatory ability due to the greater divergence in allele frequencies; second, as the allele frequencies are well-estimated for major ethnic groups, it would be possible to create pseudo-classes that have class-specific allele frequencies fixed to these values. In this way, it might become apparent, for instance, that a large proportion of a sample is an

admixture between Caucasian and African-American ancestry despite the fact that there are no pure African-American individuals in the sample. That is, currently, to detect a class as admixed, the ancestral classes must exist in the sample in the pure form also. Online resources such as the ALFRED database project should assist this effort (Cheung et al., 2000).

A final issue regards the optimal use of the posterior probabilities as a covariate. In particular, the impact of using an incorrect K solution needs to be investigated, especially if the AIC criterion has a tendency to over-estimate K.

Part III

Sample Selection
and
Complex Effects

Chapter 7

Selection & gene–environment interaction

This Chapter considers the incorporation of measured environmental moderator variables in three separate contexts relevant to selected samples: 1) modelling environmentally-moderated QTLs in variance components linkage analysis, 2) enhancing sample selection methods by the use of measured environmental moderator variables, and 3) modelling environmentally-moderated QTLs in association analysis.

7.1 Introduction

Gene–environment interaction ($G \times E$) is most tractable when dealing with measured (as opposed to latent) genetic and environmental effects. In Chapter 4, which deals with latent $G \times$ measured E interaction, an approach is outlined with reference to the classical twin study. In this Chapter the same approach is extended to quantitative trait locus (QTL) sib-pair linkage analysis, within a variance components framework. Although analysis of $G \times E$ should eventually lead to a better understanding of the aetiology of complex traits and diseases, in the context of linkage analysis (which only identifies fairly large genomic regions likely to harbour disease-causing genes)

the current goal is simply to increase power to detect genes of small effect, rather than to dissect genetic–environmental architectures *per se*.

Imagine that twin analyses have indicated that genetic influences on a trait increase with age. In planning a QTL mapping study, it might follow that the investigator should focus on older populations in order to maximise chances of detecting QTL. This may be either because the same genes have greater effects in older people, or because novel sets of genes operate in older people and not in younger people. Of course, it must be kept in mind that evidence for $G \times E$ from a twin study has no necessary implication for the genetic architecture of any one QTL. Simply because the polygenic additive effect is moderated by age, not every QTL would necessarily be expected to demonstrate increased effects in older individuals. The first section in this Chapter illustrates a method of incorporating measured environmental variables, such as age, into variance components linkage analysis. The correct modelling of a specific QTL × environment interaction could potentially increase power to detect loci of small effect.

There is a second way in which consideration of interaction effects could increase the power of QTL studies. For QTL linkage we know that, aside from the QTL variance itself, power is massively influenced by the *residual* sibling correlation. For example, if a QTL accounts for 5% of the trait variance, a test of linkage will be more powerful if most of the 95% residual variance is shared amongst siblings. Figure 7.1 plots the number of individuals required for 80% power and type I error rate of 5% for a 10% QTL when looking at a complete-information marker with a recombination fraction of 0.1 with the QTL. This quantity is calculated for three different levels of residual correlation, for unselected pairs, trios, quads and quints. As these are unselected samples the numbers required are very large, although increasing sibship size increases efficiency. However, the impact of the residual correlation is also considerable.

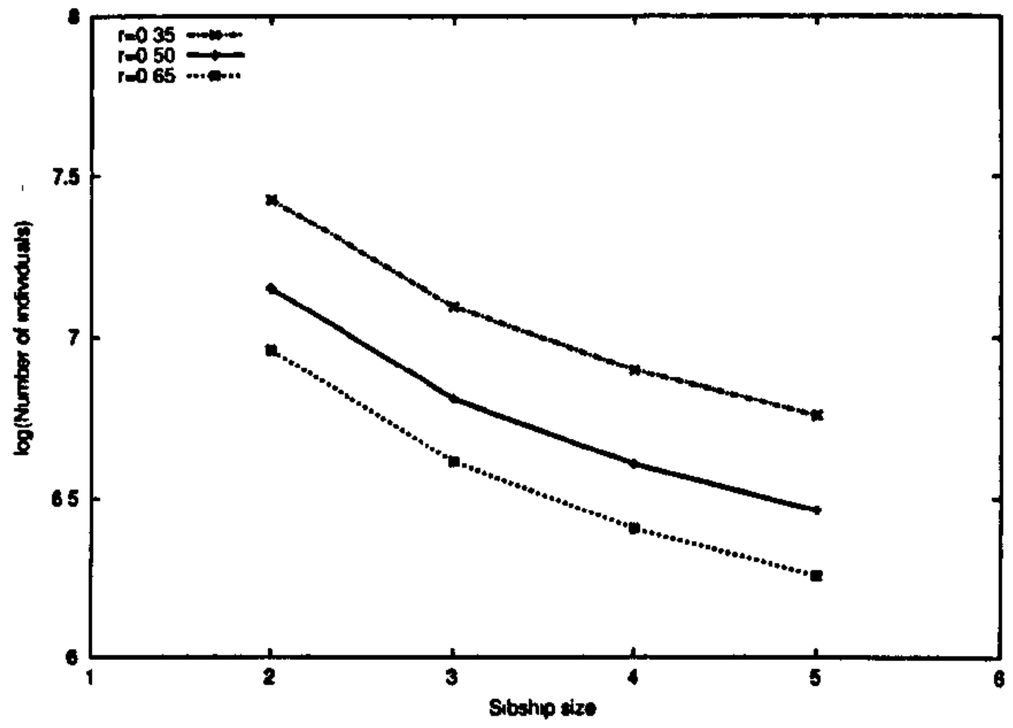


Figure 7.1: The impact of residual sibling correlation on power of QTL linkage: the log of the number of individuals required for different residual sibling correlations (QTL accounts for 10%, total sibling correlation is 0.35, 0.5 or 0.65). See text for further details.

The investigator can not normally manipulate the sibling correlation in a population in such a way as to lead to a more powerful test of linkage. However, any evidence of heterogeneity in terms of interaction effects suggests that the population actually consists of different subpopulations, which may have different expected familial correlations. Incorporating this information into a selection procedure (i.e. selecting on moderator variables) as well as selecting on extreme trait scores may increase power. The model of $G \times E$ interaction proposed in Chapter 4 also allows for "$E \times E$" interaction. For example, a specific measured E variable might interact with an anonymous, latent environmental variable. If a measured variable is found to moderate the shared environmental component of variance, such that pairs with high moderator values will also have higher expected sibling correlations, this variable could be incorporated into a selection scheme. In short, any variable which moderates the residual sibling correlation, whether this is via genetic means or not, is potentially valuable.

The third section of this Chapter extends the conditional approach to QTL association analysis presented in Chapter 3 to allow for QTL effects that interact with measured environments. This design is potentially the most powerful to detect $G \times E$.

7.2 Gene–environment interaction and QTL linkage in selected samples

The variance components approach to sib-pair QTL linkage analysis is, in essence, only a trivial extension of the twin model (Kruglyak and Lander, 1995a; Amos, 1994; Fulker et al., 1999). Assuming we have only full sibling pairs, the likelihood is parameterised in terms of three variance components: variance due to the QTL, Q, variance due to shared sibling effects, S, and variance due to nonshared sibling effects, N. Polygenic additive effects load onto both S and N. The basic allele-sharing test of linkage is of the relationship between phenotypic sib-pair similarity and IBD sharing at the test locus. The "weighted-likelihood conditioning-on-trait-values" approach Sham et al. (2000a) is adopted in the following analyses, in order to provide a robust test of linkage in selected samples.

In this section, only the scenario where the actual QTL effect is moderated by a measured covariate ($Q \times M$) is considered. The next section considers the scenario where a residual variance component is moderated by a measured covariate (i.e. $S \times M$ and $N \times M$) and shows how knowledge of this can be used to enhance sample selection strategies.

7.2.1 $Q \times M$ in linkage analysis

Analogous to the modelling of a moderating effect on the additive genetic path, a, in the twin model (Chapter 4), the QTL path q is simply modified to $(q + \beta_Q M)$ or even $(q + \beta_Q M + \delta_Q M^2)$ to incorporate $Q \times M$ interaction, representing linear and nonlinear interactions respectively, between the additive genetic value at the QTL, a_Q, and the moderator M. The presence or absence of a particular allele is assumed to be unrelated to the moderator (i.e. no gene–environment correlation).

The simulations reported in Table 7.1 show four conditions varying in (1) QTL

Simulated			Likelihood ratio tests		
			$Q - SN$	$Q - SN - X_Q$	$Q - SN - X_Q - X_S X_N$
a_Q	β_Q	β_X	SN	SN	$SN - X_S X_N$
0	.	.	0.62	1.93	1.77
0	.	0.2	0.58	13.74	1.81
2	.	.	48.10	48.27	48.67
2	0.3	.	45.13	106.87	49.66

Table 7.1: QTL linkage incorporating $Q \times M$ interaction in DZ twin pairs, with and without $A \times M$ interaction also.

effect ($a_Q > 0$) (2) $Q \times M$ interaction ($\beta_Q > 0$) and (3) residual $G \times E$ interaction ($A \times M$ in fact, i.e. $\beta_X > 0$). For each condition 200 replicate datasets are simulated, and a number of likelihood ratio test statistics constructed. The base model SN has no QTL effects; model $Q - SN$ allows for a simple QTL effect; model $Q - SN - X_Q$ allows for a moderated QTL effect as well as a main effect. Two additional models also allow for the possibility of interaction effects between the residual variance components (S and N) and the measured moderator variable M. From left to right, the three likelihood ratio tests shown in Table 7.1 therefore represent (1) a simple 1 degree of freedom test for an additive QTL effect (2) a 2 degree of freedom test for a QTL effect that might be moderated by the variable M and (3) as for the previous test, but allowing for $S \times M$ and $N \times M$ effects under both the super- and submodel. In all cases, 1000 DZ twin pairs were simulated, with residual variance components $a = c = e = 1$ and an additive diallelic QTL with equal allele frequencies. The expected variance components associated with the $Q \times M$ corresponding to the fourth row of Table 7.1 are illustrated in Figure 7.2.

The first two rows of Table 7.1 represent the case of no QTL effect. In the first row (no QTL effect, no interactions) the test statistics are all close to their expected values under the null. The second row (no QTL effect, residual interaction) shows that the combined test of a *moderated* QTL effect (second column) is highly anti-conservative in the presence of residual $A \times M$, with a highly significant $\chi^2 = 13.74$ (expected χ^2

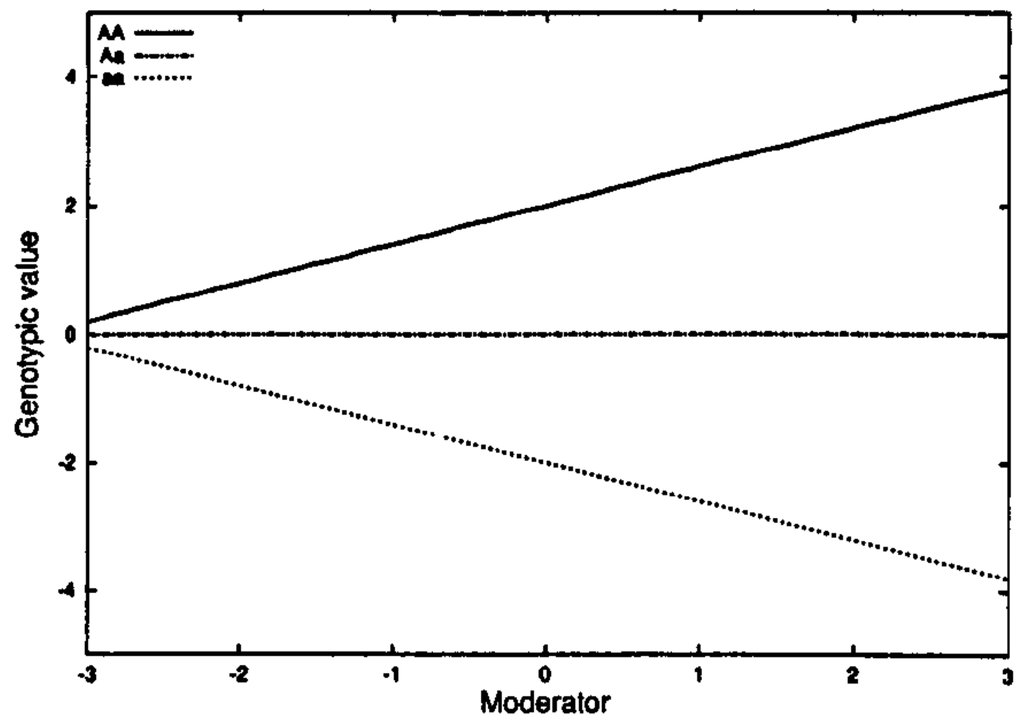


Figure 7.2: Example of a simulated $Q \times E$ interaction $\beta_Q = 0.3$ with additive genetic value $a_Q = 2$ and dominance deviation $d_Q = 0$.

is 1.5, as the test of Q only involves 0.5 degree of freedom). This bias is due to the greater variance at higher levels of the moderator (due to the residual interaction) which the β_Q parameter attempts to account for. Properly modelling this residual non-additivity (i.e. by the inclusion of $S \times M$ and $N \times M$ terms, as in the third column) reduces this bias. Therefore, it is unwise to perform a simple $Q \times M$ type of analysis when conducting a linkage test when there is non-additivity in the data.

The next two rows of Table 7.1 represent the case of a large QTL effect ($a_Q = 2$). In the third row (QTL effect, no interactions) the likelihood ratio tests are all similar (although the first has one less degree of freedom). In the fourth row (QTL effect, QTL interacts with moderator), the 'robust' linkage test (third column) gives very little extra information compared to a simple QTL test (first column). This is because the $S \times M$ and $N \times M$ components will soak up the $Q \times M$ effect, i.e. the opposite of the above effect. If one were able to be sure that there were no significant residual interaction effects, however, then the basic test of a moderated QTL effect (second column) would in fact provide more power under the alternate.

Overall, these results illustrate some of the potential gains and losses involved with modelling $G \times E$ in the QTL linkage analyses. In particular, a simple approach to $Q \times M$ interaction can lead to false positive results.

7.3 Residual interaction and sample selection for linkage

As discussed in Chapter 2, the use of selective sampling schemes for linkage is highly desirable, especially when working with sibling pairs, as most pairs will yield very little information for linkage. As mentioned, irrespective of QTL effect size, a higher residual correlation increases power to detect a QTL (Sham et al., 2000b).

Typically, a single value for the sample residual correlation is specified when selecting or analysing a sample for linkage. However, in the presence of $G \times E$ there will, by definition, be heterogeneity in the residual correlation across the sample. This section explores the possibility of using prior knowledge of such heterogeneity (when the relevant moderating variables have also been measured in the linkage sample) to better specify pair-specific residual correlations in order to increase power.

A correlation is a property of a number of paired observations: specifying a pair-specific correlation implies that the pair belongs to a particular subset which has that correlation. If a moderator variable M interacts with either A, C or E components, then M can predict which pairs will have higher residual sibling correlations. Consider, for example, an $E \times M$ interaction such that individuals scoring higher on M will tend to have lower effects of E. In this case, pairs in which both members score high on M will have a higher residual correlation. All other things being equal, it would therefore be preferable to select this pair over a pair with a lower residual correlation.

Figure 7.3 illustrates the relationship between sample selection and $G \times E$ in three graphs. Panel a) illustrates the relationship between trait score and expected informativeness: concordant high and low pairs and discordant pairs in particular are most informative. Panel a) assumes a constant sibling correlation across the sample however, which might not be the case. Panel b) illustrates how the residual sibling correlation might change as a function of a moderator variable, in the presence of an

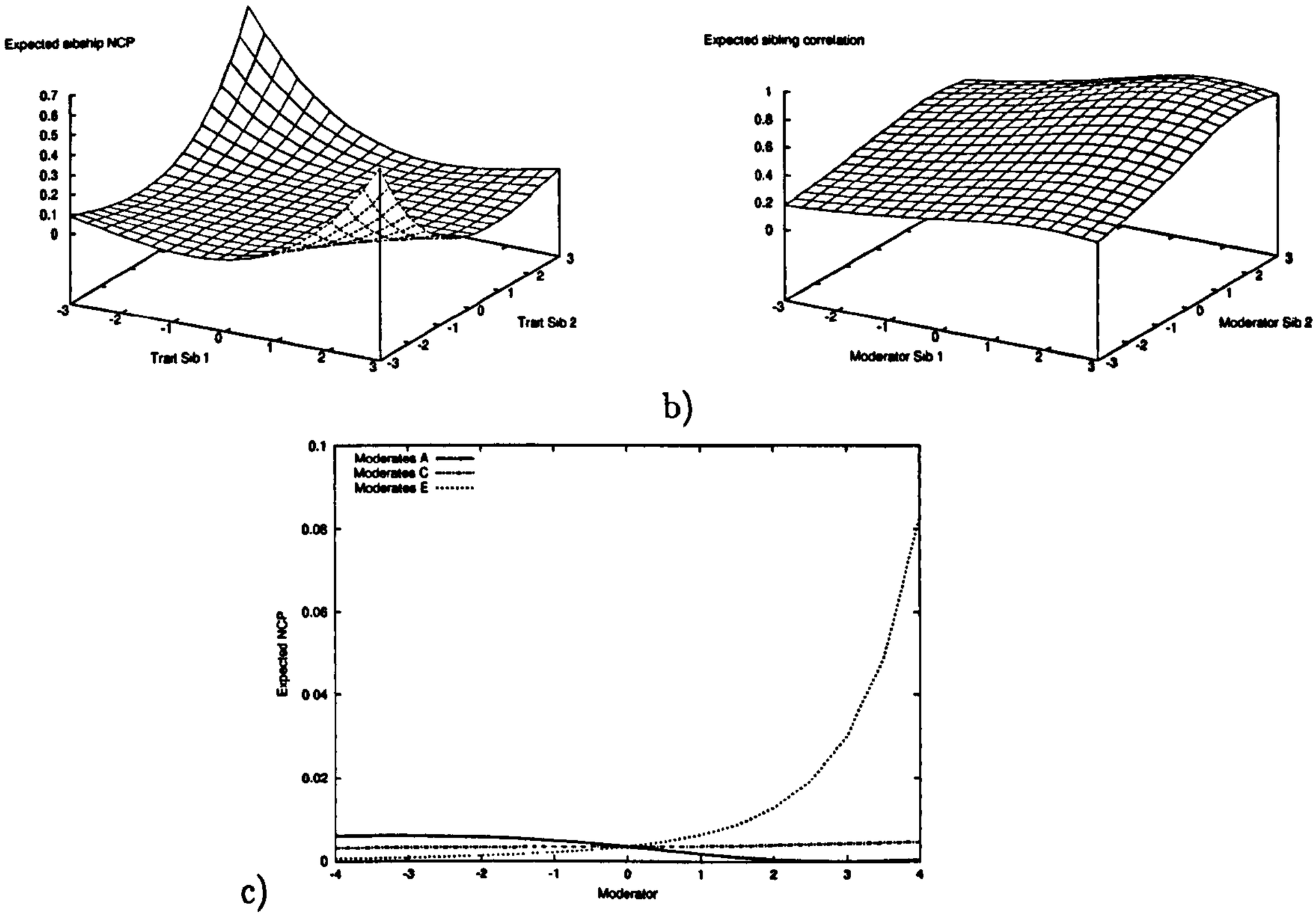


Figure 7.3: $G \times E$ and sample selection for linkage. See text for explanation.

$E \times M$ interaction similar to that described above. It would therefore be desirable to take this information into account when selecting and analysing pairs for linkage: panel c) shows the marked impact on the expected non-centrality parameter (via the expected residual correlation) for the linkage test in the presence of $G \times E$. The graph shows the expected non-centrality parameter (NCP) per randomly-selected sib pair as a function of sib pair moderator (assuming, in this case, that the moderator is identical between sibs and that the main effect of the moderator has been partialled out of the trait). In particular, modelling $E \times M$ interaction can greatly increase power – it seems that residual $A \times M$ and $C \times M$ do not influence the test so much (as they operate on both sibling variance and covariance, and so have less impact on the correlation).

It is interesting to note that these results are related to an observation regarding bivariate linkage analysis and the source of residual cross-trait phenotypic covariance: that power increases dramatically with decreasing nonshared sources of covariance (Evans, 2002). In this sense, bivariate analysis and including a moderator variable can have a similar effect: the impact on the NCP of modelling $E \times M$, as shown above in panel c), seems to reflect a similar trend to that shown in Figure 2 of Evans (2002). It seems possible, therefore, that the benefits of bivariate linkage can be harnessed within a $G \times E$ framework with a moderator that interacts with the residual nonshared component, whether or not the second trait is at all related to the test QTL.

Focusing on $E \times M$, we assume that prior twin analyses have estimated a significantly nonzero value for β_Z. For a phenotyped sample of pairs also measured on M, this prior knowledge can be used (1) to select sibling pairs which are most informative for linkage, by calculating the residual correlation applicable to that pair conditional on measured M and (2) in analysis, to use the pair-specific residual correlations. Ideally, the sample in which β_Z was estimated will be as close as possible to the linkage sample (for example, the linkage sample could be all the DZ pairs from the twin

sample). Effects of misspecifying β_Z are explored below in the simulations.

Using a method based on the Haseman-Elston linkage test (Haseman and Elston, 1972; Sham and Purcell, 2001), the expected noncentrality parameter (NCP) for pair i is

$$\frac{q^4}{16}\left[\frac{(T_{i1}+T_{i2})^2}{(1+r)^2}-\frac{(T_{i1}-T_{i2})^2}{(1-r)^2}+\frac{4r}{1-r^2}\right]^2$$

assuming complete marker informativeness, where T_{i1} and T_{i2} are the standardised trait scores for the pair, r is the sibling correlation, and q^2 is the proportion of variance due to the QTL. This index can be used to rank order sibling pairs by potential informativeness. In the presence of heterogeneity, it is possible to calculate pair-specific correlations, which will more accurately model the residual variance in the sample. For pair i, conditional on estimated values of a, c, e and β_Z and measured M_{i1} and M_{i2}, then r_i can be calculated as

$$r_i = \frac{0.5 \times a^2 + c^2}{\sqrt{a^2+c^2+(e+\beta_Z M_{i1})^2}\sqrt{a^2+c^2+(e+\beta_Z M_{i2})^2}}$$

which can be substituted in the above expression. The trait score for sib j of pair i, T_{ij}, also has to be standardised to unit variance conditional on the moderator. In the case of a, c, e and β_Z having been previously estimated

$$T'_{ij} = T_{ij}/\sqrt{a^2+c^2+(e+\beta_Z M_{ij})^2}$$

although the expressions for the moderator-conditional standardised scores and correlations will change depending on which models are being used to give the prior parameter estimates. Sibships, not twins, may only have been available, for example.

The formulation of the linkage model used here (Sham et al., 2000a) has only a single free parameter, the QTL variance, q^2. The total variance and residual correlation are fixed, either to their sample values or other values estimated in previous

studies (e.g. in the case of a selected sample). In the present case, the variance is fixed to unity and the residual correlation fixed to the pair-specific values, conditional on the moderator. The covariance matrices conditional on IBD sharing at the test locus for pair i are therefore $\begin{bmatrix} 1 & r_i - q^2/2 \\ r_i - q^2/2 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & r_i \\ r_i & 1 \end{bmatrix}$ and $\begin{bmatrix} 1 & r_i + q^2/2 \\ r_i + q^2/2 & 1 \end{bmatrix}$ for pairs sharing 0, 1 and 2 alleles IBD, respectively.

7.3.1 Simulations

Simulations based on sib-pair datasets featuring a residual $E \times M$ interaction in all cases were conducted under a number of conditions: varying QTL effect, sample selection scheme and whether or not the residual interaction was included or misspecified in the analysis model (Table 7.2). Under each condition a dataset of 5000 DZ pairs was simulated 200 times. Selected sample analyses were based on the most informative 10%, i.e. 500 pairs. The QTL effect was specified in terms of the additive genetic value, a_Q, which was 0, 0.5 or 1, for a fully informative diallelic test locus with equifrequent alleles. Three final conditions concerned the residual interaction, which was simulated as $\beta_Z = 0.5$ in all cases (illustrated in Figure 7.4). In the first case, "w/ $E \times M$", the correct moderator variable was incorporated into the analysis with the correct estimate of β_Z to form the pair-specific residual correlations used in selection and analysis. In the second condition, "w/ out $E \times M$", both selection and analysis were performed as usual, ignoring the moderator M. In the third condition, the true moderator was replaced with an unrelated random variable (i.e. which would have no moderating properties with respect to the trait) but β_Z was still assumed to be 0.5, representing a misspecification of the moderating effect in selection and analysis.

Under the null of no QTL effect ($a_Q = 0$), all models show average test statistics close the expected value (0.5), whether or not the moderator was included or misspecified and whether or not the analysis was performed on the whole or a selected

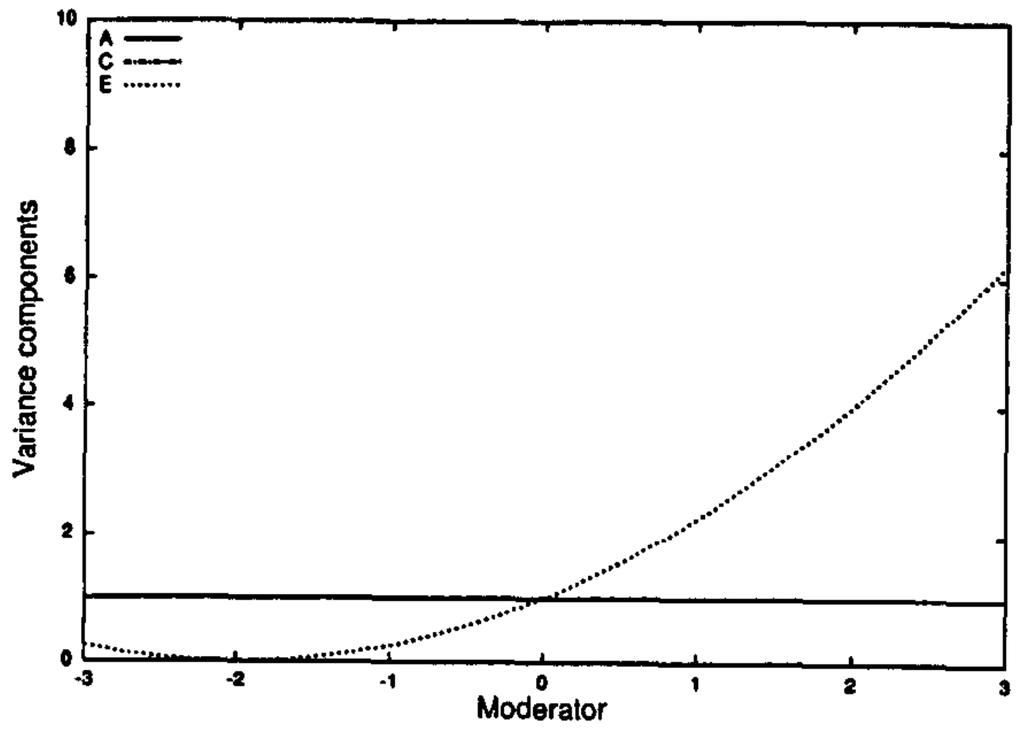


Figure 7.4: $E \times M$ interaction with residual components $a = c = e = 1$ and $\beta_Z = 0.5$, as used in all simulations.

	10% most informative					Unselected				
			% significant at $p =$					% significant at $p =$		
a_Q	$\hat{q}$	LRT	0.025	0.005	0.0005	$\hat{q}$	LRT	0.025	0.005	0.0005
	w/ $E \times M$									
0	0.012	0.53	2	1	0	0.010	0.56	2.5	1	0.5
0.5	0.047	2.73	27	9.5	2.5	0.042	3.21	30	14	4.5
1	0.180	20.18	100	98	88	0.184	31.82	100	100	99
	w/out $E \times M$									
0	0.011	0.43	1.5	0	0	0.010	0.53	2.5	1	0
0.5	0.028	1.58	14	3.5	1	0.033	2.35	24	9	1.5
1	0.101	11.40	90.5	75.5	45	0.121	21.30	99.5	95.5	91.5
	w/ incorrect $E \times M$									
0	0.006	0.57	3	0.5	0	0.005	0.43	2	0	0
0.5	0.015	1.29	9.5	2.5	0	0.016	1.70	15.5	5	1
1	0.079	9.41	80.5	62	34	0.100	17.53	97	91	78.5

Table 7.2: Results of QTL linkage simulations incorporating $E \times M$ interaction.

sample. The $\hat{q}$ column gives standardised estimates of the QTL variance, which are all close to zero under the null. For the selected and unselected samples, Table 7.2 also gives the % of replicates (out of 200) significant at various significance levels, which are all close to expected values.

Under the alternate hypothesis (i.e. $a_Q > 0$) it is clear that selected samples are more efficient than unselected samples (e.g. for $a_Q = 1$, in the condition not incorporating the moderator, on average 54% (11.40/21.30 = 0.535) of the signal was recovered by 10% of the sample). Incorporating the moderator results in a considerable gain in information. In terms of the average test statistic, for $a_Q = 1$ in unselected samples, there is a gain of 50% (i.e. comparing "w/ $E \times M$" and "w/ out $E \times M$" conditions, (31.82-21.30)/21.30). For $a_Q = 1$ in selected samples, there is a gain of 77% ((20.18-11.40)/11.40). In terms of the percentage significant with this sample size at a particular significance level, the gains can be great; for example, 88% are significant for $a_Q = 1$ at $p = 0.0005$ when the moderator is included compared to only 45% when it is not.

The "w/ incorrect $E \times M$" rows represent the scenario where the moderator is actually completely unrelated to the trait (i.e. the estimate of β_Z obtained from another dataset is completely unwarranted in this one). As can be seen, this does reduce power to some extent, although the test still appears to have the correct performance under the null. In the case of $a_Q = 1$ the average test statistic drops by approximately 18% for both selected and unselected samples, the majority of the signal remains intact despite the complete misspecification.

If there is strong reason to believe that the moderating effect does exist in the linkage sample, then both selecting and analysing incorporating the moderator seems desirable. If the effect is less certain, then it might not be advisable to select on the basis of the moderator, although it would be of interest to conduct the analyses both with and without incorporation of the putative moderator.

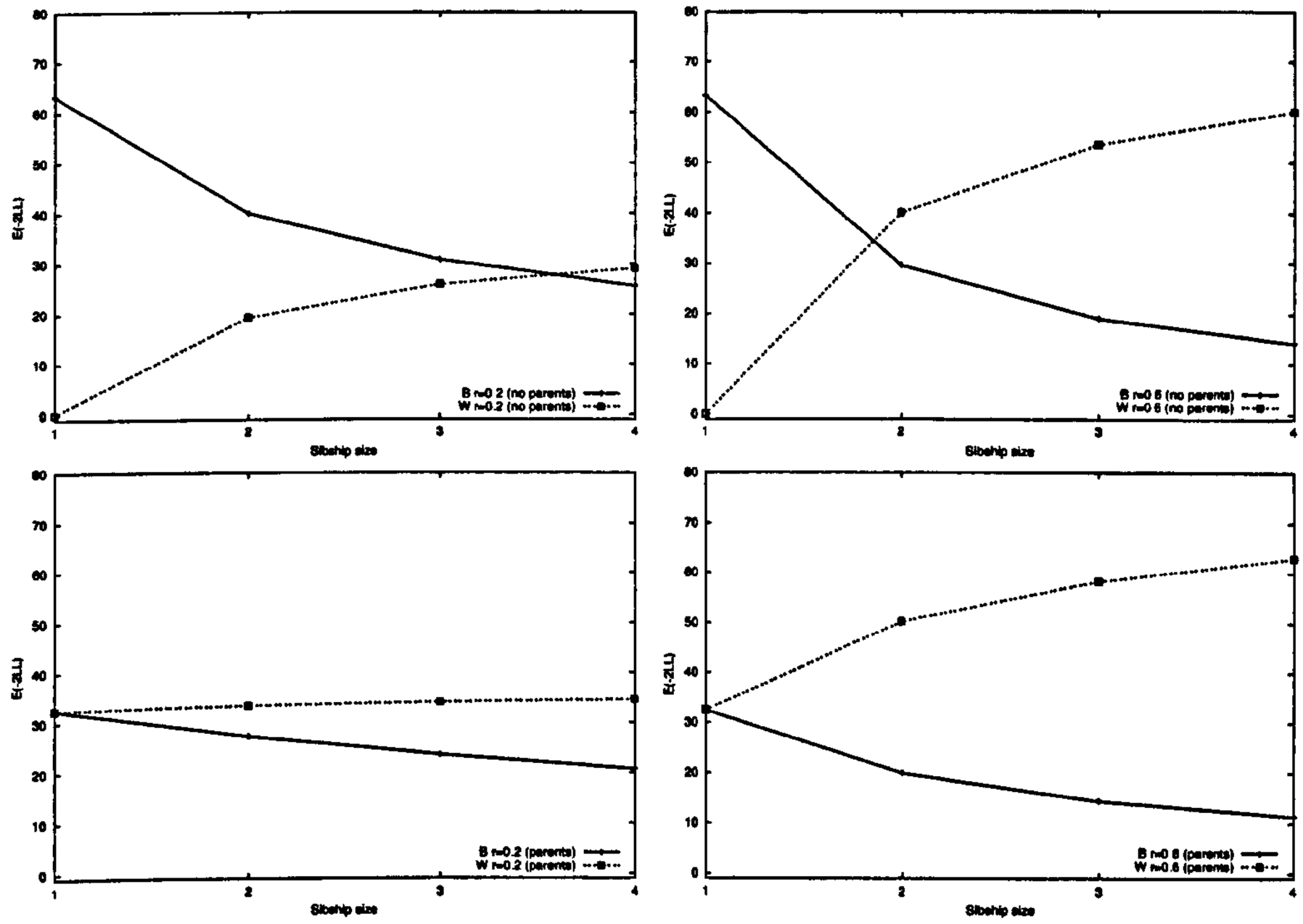


Figure 7.5: Impact of residual sibling correlation on between and within components of association in sibships, size 1 to 4. The top row of graphs represent the case of no parental genotypes; the bottom row when parental genotypes are available. The left column represents a residual sibling correlation $r = 0.2$; the right column represents $r = 0.6$.

7.3.2 Selection for QTL association

The power of family-based association analysis also depends on the residual sibling correlation: the relative balance of between and within sibship components is particularly influenced. For example, consider an additive QTL accounting for 10% of the trait variance. For six hundred individuals, either as 300 pairs, 200 trios or 150 quads, the sample NCPs calculated under a residual correlation of 0.2 and 0.6 for the between and within components of association are shown in Figure 7.5. The power of the robust within-sibship test increases with both increased sibship size and increased residual correlation. It may also be desirable to extend the approach described in the previous section to association, i.e. using sibship-specific correlations calculated conditional on a moderator variable for selection and analysis.

7.4 Gene–environment interaction and QTL association in selected samples

In this final section, the approach described in Chapter 3 for association analysis in samples of selected sibships is extended to incorporate gene–environment interaction. The environmental moderator may be a binary or continuous variable, which is assumed to moderate the QTL effects in a linear manner.

Adopting the conditional approach, modelling the likelihood of observing genotype conditional on trait, an environmental moderator variable represents an extra factor to condition on. That is,

$$L(G|X,E) = \frac{L(X|G,E)L(G|E)}{\sum_G L(X|G,E)L(G|E)}$$

although, for the time being, we assume that G and E are independent in the population (i.e. no gene–environment correlation, r_{GE}), so

$$L(G|X,E) = \frac{L(X|G,E)L(G)}{\sum_G L(X|G,E)L(G)}.$$

The biometrical model used to specify $L(X|G,E)$ is outlined below, followed by simulation results.

7.4.1 Biometrical model

The basic biometrical model (Falconer, 1989) describes the three genotypic means of a diallelic QTL as $m+a$, $m+d$ and $m-a$ for genotypes 1/1, 1/2 and 2/2 respectively. To incorporate linear $G \times E$, three new parameters are needed: β_M, β_A and β_D. Conditional on a measured moderator variable M, the genotypic means are now

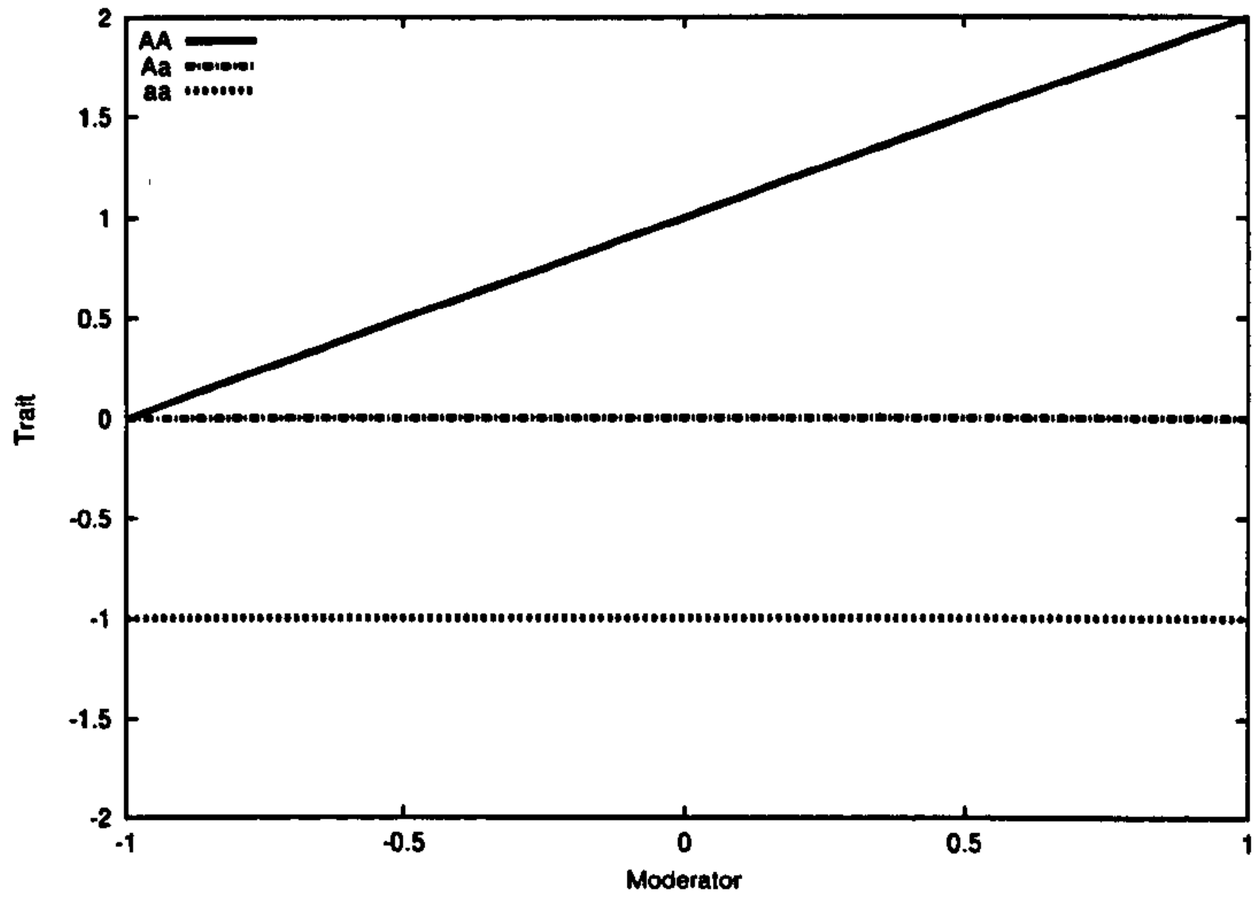


Figure 7.6: Illustration of $G \times E$ interaction for a specific QTL: see text for further details.

Genotype	Expected value
1/1	$m + \beta_M M + (a + \beta_A M)$
1/2	$m + \beta_M M + (d + \beta_D M)$
2/2	$m + \beta_M M - (a + \beta_A M)$

The mean parameter m is fixed to zero, as the data are assumed to be mean-centred prior to analysis (using the population mean in the case of selected samples). As illustrated below, the expected mean vectors are also mean-centred. The moderator variable must also be standardised, and the population sibling correlation for the moderator must be specified also.

The parameters β_M, β_A and β_D allow the genotypic means to vary as a function of the moderator variable. For example, consider the situation where there is an additive main effect for a QTL, except one of the homozygotes interacts with the moderator: Figure 7.6 illustrates this scenario. This example actually involves all three types of interaction. In this case, the parameter values are $a = 1$, $d = 0$, $\beta_M = 0.5$, $\beta_A = 0.5$ and $\beta_D = -0.5$. As the mean of the moderator is 0 and the interactions are linear, it is clear that the action of the QTL averaged over all environments in additive: the heterozygote score is exactly halfway between the two homozygotes. Conditional on

the moderator, the QTL operates in a partially dominant manner as one moves away from the moderator mean, i.e. there is an interaction between the moderator and dominance effects of the QTL, so $\beta_D \neq 0$. The distance between the two homozygotes changes as a function of the moderator, so there is also an interaction between the moderator and additive effects of the QTL, so $\beta_A \neq 0$. Finally, conditional on the moderator, the overall mean changes: this represents a main effect of the moderator on the trait, so $\beta_M \neq 0$.

A test of additive $G \times E$ is therefore $H_A(a, \beta_M, \beta_A)$ against $H_0(a, \beta_M, \beta_A = 0)$. That is, β_M is still free under the null model, so the test is a 1 degree of freedom test. More than one moderator can be included by simply adding more interaction terms in the means model. When this approach is implemented within a conditioning-on-trait-values framework, some further issues arise. Because the likelihood now involves the probability of observing the genotype conditional on the trait, the main effect of the moderator on the trait has no direct impact on the likelihood. A model that attempts to estimate only β_M without allowing for a main effect of QTL will necessarily result in a χ^2 test of 0. If there is a main effect of genotype, and it is estimated, then it is possible to estimate β_M, although power to detect it within the conditional framework will be low. As illustrated above, the β_M parameter is necessary in order to be able to describe the entire range of possible $G \times E$ models, however. This leads to the question of whether the above additive $G \times E$ test is exactly a 1 degree of freedom test. If β_M is not identified under the null, then it could be argued that it should be fixed too, giving a 2 degree of freedom test. This is analogous to the dependence of allele frequency and additive genetic effect in complex segregation analysis, e.g. when $a = 0$ then p is empirically under-identified. The simulation results presented below address this issue, by considering type I error rates for the different degrees of freedom.

Using the conditional approach, the calculation of the expected trait values and residual covariance matrix for each genotypic configuration (GC) is now conditional on the sibship's scores on the moderator. Rather than being calculated in advance of iterating over each sibship, they must be calculated anew for each sibship (unavoidably slowing analysis). For each sibship, for each possible GC, for each individual i, the expected trait score conditional on that individual's moderator M_i is

$$\begin{aligned}[\mu]_i &= (a_b + \beta_A M_i)[\mathbf{A_b}]_i + (a_w + \beta_A M_i)[\mathbf{A_w}]_i + \\ &\quad +(d_b + \beta_D M_i)[\mathbf{D_b}]_i + (d_w + \beta_D M_i)[\mathbf{D_w}]_i + \\ &\quad +\beta_M M_i - a_b(p-q) + 2pqd_b\end{aligned}$$

where A_b, etc, are defined in Chapter 3 (if parental genotypes are available, A_{bp} is used instead, etc). Note that the same interaction parameters are used for the between and within components. Although it is possible to estimate different interaction parameters for between and within effects, significant differences would be hard to interpret. In the simulations below, the condition $a_b = a_w$ always holds. If there is evidence of stratification in the sample, then any further analyses should probably only focus on the within component of association. Otherwise, when incorporating $G \times E$, the total component of association (i.e. $a_b = a_w$) should be used.

The expected residual variance is calculated conditional on the moderator also: for each sib i, the shared and nonshared residual variances are

$$\begin{aligned}\sigma^2_{Si} &= r - \frac{\sigma^2_{Ai}}{2} - \frac{\sigma^2_{Di}}{4} - r_M\beta^2_M \\ \sigma^2_{Ni} &= (1-r) - \frac{\sigma^2_{Ai}}{2} - \frac{3\sigma^2_{Di}}{4} - (1-r_M)\beta^2_M\end{aligned}$$

where r is the fixed sibling trait correlation and σ^2_{Ai} and σ^2_{Di} are the variances explained by the QTL for sib i, conditional on M_i, and r_M is the fixed sibling moderator cor-

relation. The additive and dominance QTL variances, conditional on the moderator, are

$$\begin{aligned}\sigma^2_{Ai} &= \frac{s+1}{2s}2pq((a_b+\beta_A M_i)+(d_b+\beta_D M_i)(q-p))^2 \\ &\quad +\frac{s-1}{2s}2pq((a_w+\beta_A M_i)+(d_w+\beta_D M_i)(q-p))^2 \\ \sigma^2_{Di} &= \frac{s+3}{4s}(2pq(d_b+\beta_D M_i))^2+\frac{3s-3}{4s}(2pq(d_w+\beta_D M_i))^2\end{aligned}$$

assuming no parents (otherwise, this equation is modified as illustrated in Chapter 3). The residual variance components are bounded within [0,1] (although σ^2_{Ni} is constrained to be > 0). The elements of the residual sibship covariance matrix for sibling pair i, j are then

$$[\Sigma]_{ij} = \begin{cases} \sigma^2_{Si}+\sigma^2_{Ni} & \text{for } i = j \\ \sigma_{Si}\sigma_{Sj} & \text{for } i \neq j. \end{cases}$$

Implementation

The $G \times E$ association model is implemented within the `cafe` computer program. The following additional letters are used to construct models under the null and alternate: `c`, `g` and `G` which represent β_M, β_A and β_D respectively. If a moderator or covariate is included, the a `.cov` file is also necessary, which for each moderator specifies if β_M is to be fixed to some pre-specified value rather than estimated, and the moderator's sibling correlation.

7.4.2 Simulation results

Seven conditions were simulated, variously reflecting a main of the QTL and the moderator as well as interaction effects. For each condition, 1000 replicate samples were generated, half of which were singleton samples containing 500 individuals, the other half of which were sibling pair samples also containing 500 individuals (250

pairs). As well as the full samples, analyses were also conducted on selected samples of 100 individuals (50 pairs), based either extreme high / low groups for individuals, or maximal discordance for sibling pairs. In all cases, a diallelic QTL was simulated with equal allele frequencies, and an additive main effect accounting for 2% of the trait variance. The residual trait variance was equally split between shared and nonshared components.

In all cases, the moderator variable was simulated with a sibling correlation of $r_M = 0.5$. The 7 conditions varied in the specification of β_M, β_A and β_D, as shown in Table 7.3. The conditions are as follows: no covariate or interactive effects; covariate effect only; additive interactive effect only; covariate and additive interactive effects; covariate and additive interactive effects in different directions; dominance interactive effect only; covariate and dominance interactive effects.

Figure 7.7 illustrates the $G \times E$ effect in a single dataset simulated under the fourth model (i.e. fourth row of Table 7.3). The plot is a conditioning plot – the relationship between genotype and trait is plotted for six different, overlapping intervals of the moderator variable. Note how the magnitude and direction of effect, as represented by the loess function line, changes with respect to the moderator. This matches the pattern shown in Figure 7.8, which plots the expected and estimated genotypic means as a function of the moderator, averaged over all replicates.

Table 7.3 shows the average likelihood ratio test statistics for four different tests: of additive main effect (`--alt f --null 0`) labelled $a = 0$; of a covariate-like effect only (`--alt fc --null f`) labelled $\beta_M = 0$; of an additive interaction between QTL and moderator (`--alt fcg --null fc`) labelled $\beta_A = 0$; of additive and dominance interactions between QTL and moderator (`--alt fFcgG --null fFc`) labelled $\beta_A = \beta_D = 0$.

The results are fundamentally similar for both singletons and pairs; additionally, the pattern of results is similar for both full and selected samples, although selected

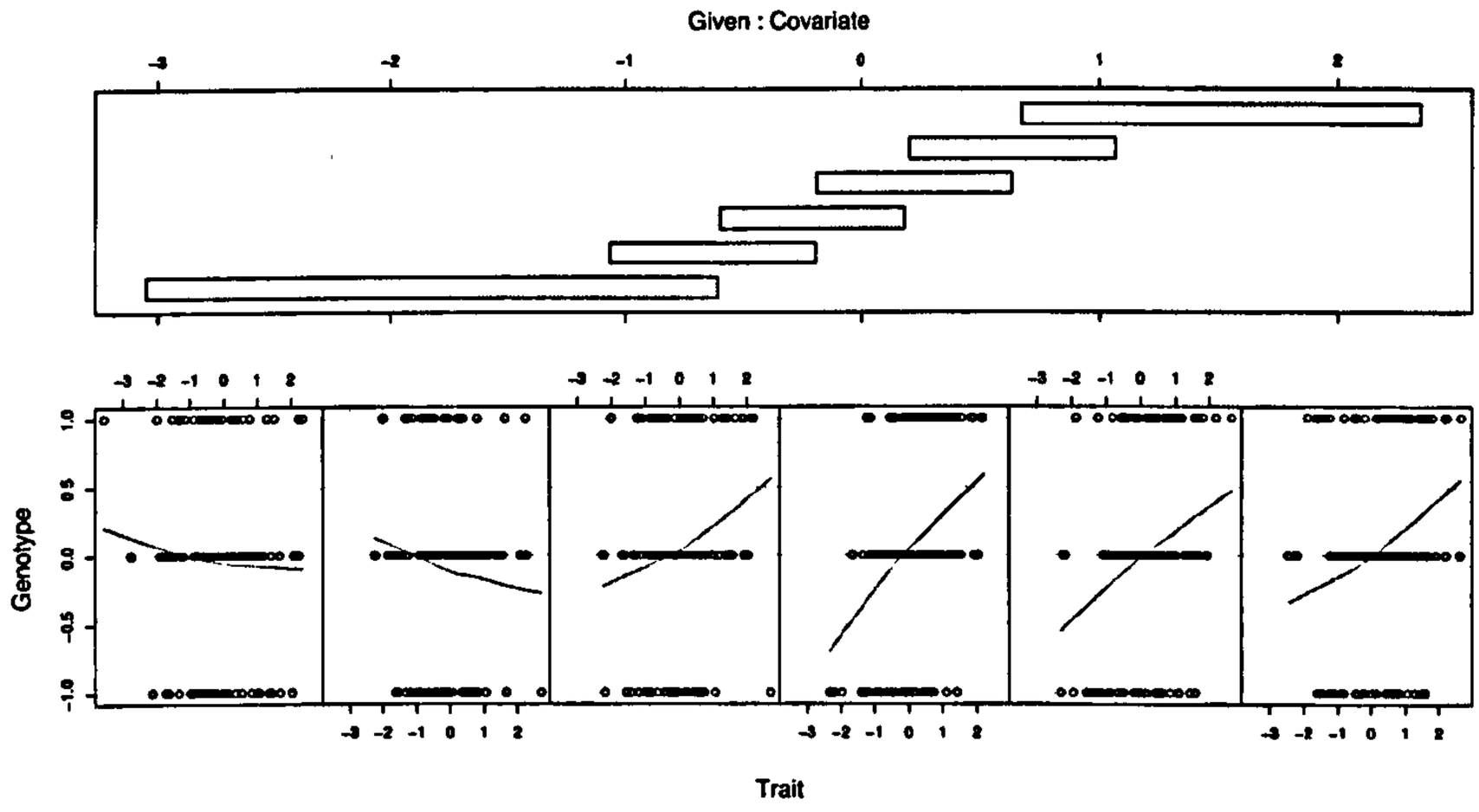


Figure 7.7: Conditioning plot of genotype on trait conditional on the moderator variable, with loess function superimposed. Corresponds to a single data set simulated under conditions specified in the fourth row of Table 7.3. Note how the direction of effect switches just below the covariate mean, as Figure 7.8 shows (fourth row).

samples show an attenuation in the test statistic when an effect is present. For the first test of a main effect of the QTL, $a = 0$, the average test statistic is equivalent across different conditions. Table 7.4 gives the power of the tests, in this case a 1 degree of freedom test with $\alpha = 0.05$. In full samples the power is around 90%, in selected samples the power is around 70% (60% for pairs).

None of the specific tests of a covariate effect, $\beta_M = 0$, show large test statistics, however, due to the nature of the conditioning on trait values approach, as suggested above. This fact is not particularly important, though. As shown in Figure 7.8, the estimates are unbiased, and the models recover the average covariate effect well. A test for a main effect of a covariate on the trait could be performed prior to the association analysis. Additionally, the association analysis could be performed on residuals after the effects of the covariate have been partialled out.

The specific tests of additive $G \times E$, $\beta_A = 0$, show good power and specificity. For conditions when additive $G \times E$ is present (i.e. rows 3, 4 and 5) the average test statistic is of the same order of magnitude as for $a = 0$. As mentioned, there is an

			$a = 0$		$\beta_M = 0$		$\beta_A = 0$		$\beta_A = \beta_D = 0$	
β_M	β_A	β_D	Full	Sel	Full	Sel	Full	Sel	Full	Sel
Singletons										
0	0	0	11.23	7.44	1.00	0.90	1.13	1.19	2.42	2.72
0.2	0	0	10.68	7.19	1.35	1.01	1.24	1.35	2.42	2.91
0	0.2	0	10.83	7.34	1.02	0.98	10.65	7.29	11.80	8.74
0.2	0.2	0	10.51	7.08	1.46	1.24	10.61	6.21	11.70	7.56
0.2	-0.2	0	10.98	7.17	1.36	1.16	10.60	6.36	11.65	7.65
0	0	0.2	11.13	7.47	1.06	1.06	1.11	1.30	7.22	5.63
-0.2	0	0.2	11.30	7.77	1.33	0.97	1.10	1.06	7.63	5.70
Pairs										
0	0	0	11.25	5.56	0.98	0.98	1.13	1.23	2.34	2.56
0.2	0	0	10.77	5.06	1.54	1.06	1.08	1.27	2.40	2.88
0	0.2	0	10.78	5.38	1.03	1.11	11.54	6.15	12.43	7.43
0.2	0.2	0	11.16	5.56	1.50	1.16	11.81	5.79	12.79	6.90
0.2	-0.2	0	10.87	5.18	1.48	1.24	12.73	6.47	13.81	7.69
0	0	0.2	11.41	5.53	0.94	1.02	1.15	1.22	8.42	5.90
-0.2	0	0.2	11.12	5.42	1.43	1.04	1.13	1.30	8.27	5.77

Table 7.3: Average test statistics for a test of $G \times E$ within the QTL association model. See text for further details.

issue regarding the appropriate number of degrees of freedom for this test. Table 7.5 gives the power for the two $G \times E$ tests assuming an extra degree of freedom (i.e. making the tests have 2 and 3 degrees of freedom respectively). Contrasting results, it seems that the true degrees of freedom is somewhere in between, and is probably a complex function of the true parameter values. There is some suggestion that treating the $G \times E$ tests as 1 degree of freedom tests (or 2 degrees of freedom if dominance interactions are also tested) is slightly anti-conservative. In practice, it is probably best to treat the tests as two degree of freedom tests, i.e. by specifying `--alt fcG --null f` (or, for dominance, as a 3 degree of freedom test by specifying `--alt fFcgG --null fF`) so that β_M is fixed under the null.

The power of the tests, even with the extra degree of freedom, seems reasonable for the full sample cases, around 70–80%. Power for the selected samples is typically around 40–50% to detect $G \times E$ interaction. The extent to which different selection strategies and different types of interaction differentially influence power is uncertain, however.

			$a=0$		$\beta_M=0$		$\beta_A=0$		$\beta_A=\beta_D=0$	
β_M	β_A	β_D	Full	Sel	Full	Sel	Full	Sel	Full	Sel
Singletons										
0	0	0	0.91	0.72	0.04	0.04	0.08	0.06	0.08	0.08
0.2	0	0	0.89	0.73	0.08	0.03	0.07	0.10	0.08	0.13
0	0.2	0	0.87	0.68	0.04	0.05	0.85	0.68	0.78	0.61
0.2	0.2	0	0.88	0.71	0.09	0.06	0.86	0.62	0.81	0.58
0.2	-0.2	0	0.88	0.71	0.09	0.06	0.88	0.65	0.81	0.56
0	0	0.2	0.89	0.74	0.05	0.07	0.06	0.08	0.52	0.40
-0.2	0	0.2	0.89	0.70	0.08	0.04	0.07	0.07	0.55	0.40
Pairs										
0	0	0	0.88	0.57	0.06	0.04	0.07	0.06	0.07	0.08
0.2	0	0	0.88	0.51	0.11	0.06	0.06	0.07	0.09	0.11
0	0.2	0	0.86	0.55	0.05	0.07	0.86	0.60	0.80	0.52
0.2	0.2	0	0.90	0.57	0.11	0.05	0.90	0.59	0.84	0.51
0.2	-0.2	0	0.89	0.54	0.10	0.09	0.93	0.62	0.87	0.56
0	0	0.2	0.89	0.57	0.05	0.05	0.06	0.08	0.63	0.44
-0.2	0	0.2	0.89	0.55	0.09	0.06	0.06	0.08	0.59	0.40

Table 7.4: Power of the main effects ($a = 0$), covariate ($\beta_M = 0$) and $G \times E$ tests, assuming 1 degree of freedom for test of β_A, and 2 degrees of freedom for the test of β_A and β_D. See text for further details.

			$\beta_A=0$		$\beta_A=\beta_D=0$	
β_M	β_A	β_D	Full	Sel	Full	Sel
Singletons						
0	0	0	0.02	0.02	0.04	0.04
0.2	0	0	0.02	0.03	0.03	0.06
0	0.2	0	0.73	0.51	0.68	0.49
0.2	0.2	0	0.74	0.44	0.70	0.40
0.2	-0.2	0	0.77	0.46	0.71	0.41
0	0	0.2	0.02	0.03	0.36	0.27
-0.2	0	0.2	0.02	0.02	0.42	0.27
Pairs						
0	0	0	0.03	0.02	0.03	0.04
0.2	0	0	0.01	0.02	0.02	0.06
0	0.2	0	0.76	0.46	0.70	0.41
0.2	0.2	0	0.78	0.40	0.73	0.35
0.2	-0.2	0	0.83	0.44	0.79	0.42
0	0	0.2	0.02	0.02	0.49	0.29
-0.2	0	0.2	0.03	0.03	0.45	0.26

Table 7.5: Power of the $G \times E$ test assuming 2 degrees of freedom for test of β_A, and 3 degrees of freedom for the test of β_A and β_D. See text for further details.

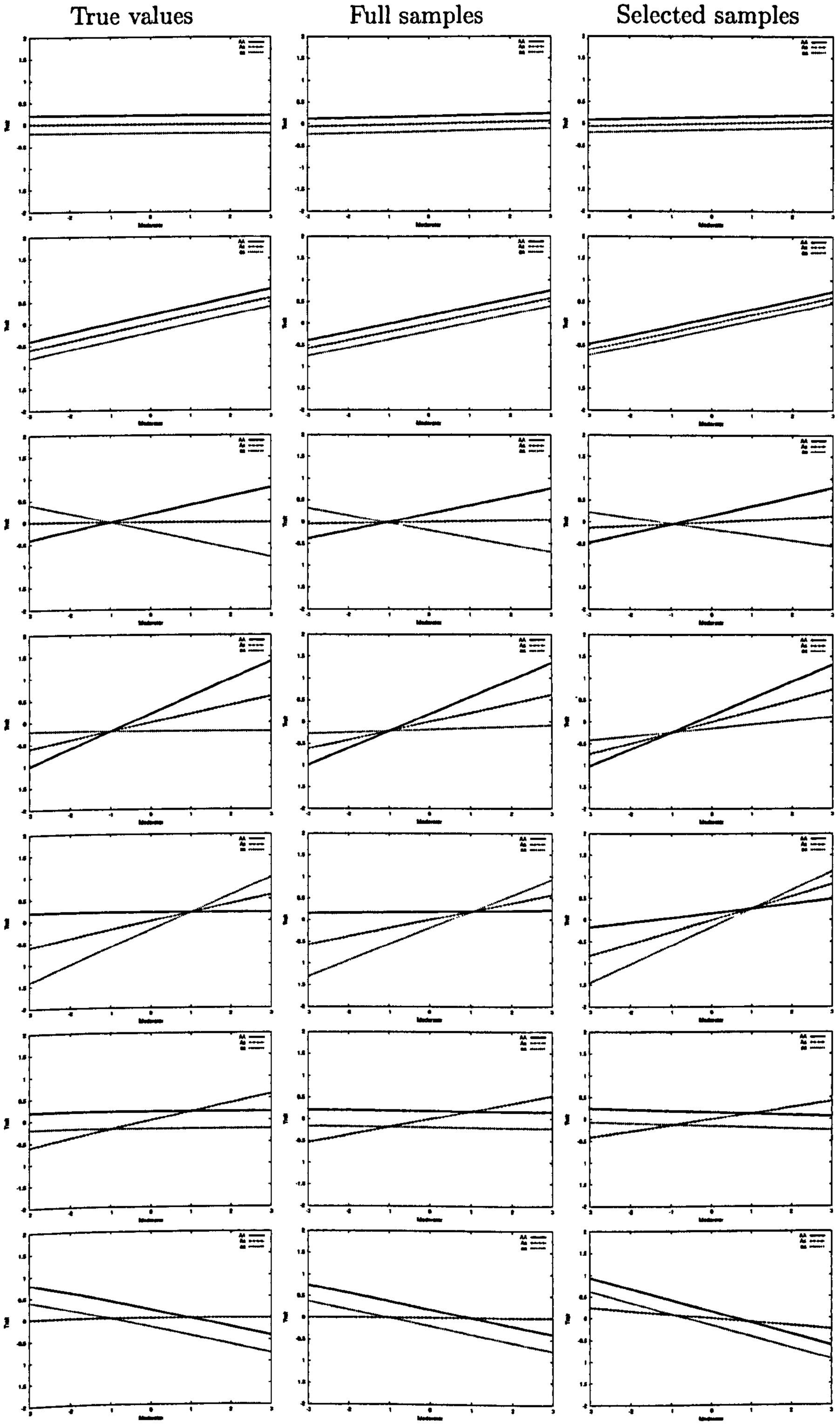


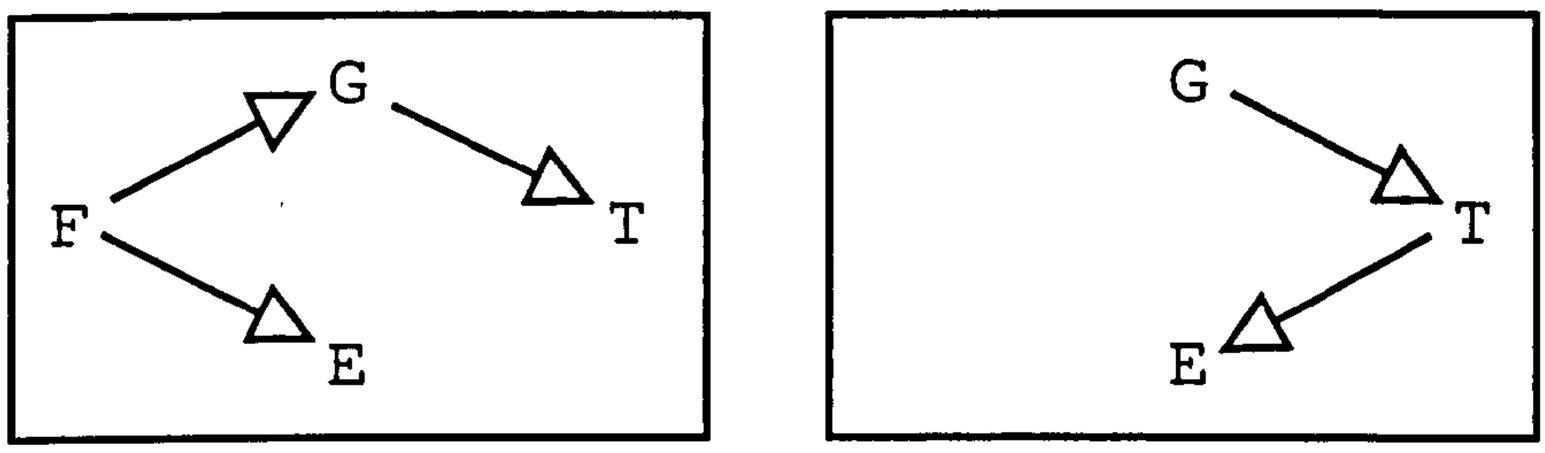
Figure 7.8: Recovery of $G \times E$ interaction in QTL association model: see text for details.

Figure 7.8 shows the recovery of $G \times E$ interaction parameters for singletons only – similar results are observed in pairs. The left column contains the expected genotypic means plotted against for the moderator, i.e. the true simulated model, for the 7 conditions, i.e. each row is one condition. The next two columns represent the average simulation results from the full and selected samples, respectively. That is, given values of a, d, β_M, β_A and β_D it is easy to plot the expected genotypic scores as a function of the moderator M. In all cases the true structure is recaptured without any noticeable bias, although there is a slight hint of the effect sizes being attenuated in selected samples.

7.4.3 Gene–environment correlation

In the conditional approach, the QTL allele frequency can be a free parameter to be estimated from the data. As mentioned above, the assumption of independence, $P(G|E) = P(G)$, was made between genotype and environment, implying no gene–environment correlation ($r_{GE} = 0$). In the presence of r_{GE}, then allele frequency would vary as a function of the environment, so $P(G|E) \neq P(G)$. Future directions for the development of this model include incorporating the ability to estimate and test for r_{GE}, as well as the impact of ignoring it. It would be possible to redefine allele frequency as a linear function of the moderator, e.g. $p + \beta_P M$, although, being a frequency, a logistic function might be rather more appropriate.

Some further issues arise when implementing this for family data, however. An intuitive view would suggest that to simply model each siblings' allele frequency as a function of their own moderator value would ignore the allelic dependence found within sibships. Several conceptual issues surrounding the notion of r_{GE} (Plomin et al., 1977) come into play in this instance, however. Passive r_{GE} implies that children born into certain environments are also more likely to have particular trait-influencing genes. Evocative and active r_{GE} imply that the individual child's trait



Passive Evocative, Active

Figure 7.9: Models of gene–environment correlation. The factors represent familial influence 'F', an individual's genes 'G', environment 'E' and trait 'T'. It is also possible that a path from E to T exists in both cases, i.e. a main effect of the environment on the trait.

value influences exposure to particular environments.

If one assumes that the r_{GE} is passive, then one should probably employ obligatory shared moderator variables. That is, otherwise it would be unrealistic to assume that two siblings from the same family who are discordant for the moderator come from populations with drastically different allele frequencies, as they come from the same parents. That is, in passive r_{GE} the source of environmental influence is via one's parents and siblings – these sources are clearly shared for siblings. However, *if* one assumes that the r_{GE} is evocative or active, then it does make sense to use moderator variables that can be different for different siblings within a sibship. In this case, the causal arrow is reversed, such that the association between gene and environment is via the individuals trait, rather than via familial factors. Figure 7.9 represents these alternatives. Note that population stratification can accurately be described as passive r_{GE}, if the environmental variable also has a main effect on the trait. A within-sibship analysis would not be able to detect passive r_{GE} therefore, even if the environmental variable wasn't obligatorily shared between siblings.

Roughly speaking, if obligatory shared moderators are employed (e.g. parental SES) any genotype–environment association would reflect passive r_{GE}. Because this analysis is obviously not possible within-sibships, it would be hard to rule out population stratification effects without further data (e.g. genomic control methods as discussed in Chapter 6). If a non-obligatorily shared moderator is used, it should

be only in a within-sibship context, to avoid dependencies in genotypes arising from the common parental genotype. In this case, any relationship between genotype and environment does not reflect different origins, but rather different consequences of different siblings within the same family having different trait values. This is likely to prove a difficult but fruitful area of future research in any case.

7.5 Summary

This Chapter has illustrated three ways in which potential measured environmental moderator variables can be incorporated into QTL analysis. For both linkage and association, interactions between a test locus and a measured environment can be incorporated; additionally, correctly modelling interactions between residual components of variance and a measured moderator can potentially increase power.

Chapter 8

Selection & epistasis

This Chapter considers approaches to two-locus models incorporating epistasis in the context of variance components QTL linkage and association mapping, with particular emphasis upon performance in phenotypically selected samples. The first method is an extension of Haseman-Elston linkage analysis to the two-locus case, based on a reformulation of the original Haseman-Elston method. The second method incorporates a second locus as a moderator variable in the association model described in Chapter 3.

8.1 Two-locus linkage analysis

As demonstrated in Chapter 5, standard variance components linkage analysis on unselected samples has very little power to detect epistatic effects formally although epistatic effects may actually contribute to the additive QTL component of variance. Chapter 2 illustrated the impact of sample selection on the efficiency of the standard linkage test – what effect, if any, sample selection has on the ability to detect epistasis is addressed in this Chapter.

This section presents two approaches to two-locus linkage analysis: one based on variance components methodology and one based on the Haseman-Elston regression

(Haseman and Elston, 1972). Simulation results are only presented for the latter method, however – as explained below, the variance components approach to two-locus linkage analysis proved to be computationally unwieldy.

8.1.1 Variance components model of two-locus linkage

The conditioning-on-trait-values approach for single-locus linkage (Sham et al., 2000a) can in theory be extended to the two-locus case, to provide a two-locus linkage test applicable to selected samples. For the additive, single-locus case, the expected sibling covariance matrix has elements for sibs i and j

$$[\Sigma]_{ij} = \begin{cases} V & i = j \\ rV + \sigma_A^2(\pi - E(\pi)) & i \neq j \end{cases}$$

where V is fixed to the population variance, r is fixed to the population sibling correlation and σ_A^2 is the additive QTL variance and single free parameter. The IBD variable, π, is the proportion of alleles shared IBD between sibs, i.e. 0, 0.5 or 1.

The two-locus case is a simple extension of this model: if only additive main and epistatic effects are considered, the expected covariance matrix is now

$$[\Sigma]_{ij} = \begin{cases} V & i = j \\ rV + \sigma_{A1}^2(\pi_1 - E(\pi_1)) + \sigma_{A2}^2(\pi_2 - E(\pi_2)) + \sigma_{AA}^2(\pi_1\pi_2 - E(\pi_1\pi_2)) & i \neq j \end{cases}$$

although dominance effects (and higher order epistatic terms) can simply be included in the covariance term, e.g. dominance × dominance interaction can be modelled by including $\sigma_{DD}^2(z_1z_2 - E(z_1z_2))$ in Σ when $i \neq j$.

For the single-locus case, the standard likelihood is a weighted sum of 3 likelihoods, corresponding to the 3 possible IBD values for a sibling pair, where the weights are based on the posterior probabilities of IBD sharing conditional on the observed marker data. The conditional likelihood has similar terms for the numerator and denominator,

except that the weights in the numerator are determined by the observed marker data, whereas the weights in the denominator are the prior probabilities of that IBD value. The conditional weighted likelihood is constructed in an analogous manner for the two-locus case, except there are now 9 possible IBD configurations. For unlinked loci the prior probabilities can be written as the matrix

$$\mathbf{P} = \begin{bmatrix} \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \\ \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \frac{1}{16} & \frac{1}{8} & \frac{1}{16} \end{bmatrix}$$

as shown in Chapter 5. Based on the observed marker data, the posterior IBD probabilities for any one sibling pair for the two loci (say 'A' and 'B') are

$$\mathbf{Q} = \begin{bmatrix} q_0^A q_0^B & q_0^A q_1^B & q_0^A q_2^B \\ q_1^A q_0^B & q_1^A q_1^B & q_1^A q_2^B \\ q_2^A q_0^B & q_2^A q_1^B & q_2^A q_2^B \end{bmatrix}$$

where, for example, q_0^A is the posterior probability of IBD 0 for locus A. The 8 IBD sharing variables described in Chapter 5 can be easily calculated from these probabilities: e.g. $\pi^A = 0.5q_1^A + q_2^A$, etc.

Each of the 9 possible IBD configurations (indexed ij where i and j take the values 0, 1 and 2) implies a particular sibling covariance matrix, $\mathbf{\Sigma_{ij}}$, and therefore a particular likelihood, given the vector of trait scores $\mathbf{y}$, for any one sibling pair (where the trait scores are mean-centred using the population mean) of

$$\ln L_{ij} = -\frac{1}{2}\left[\ln|\mathbf{\Sigma_{ij}}| + \mathbf{y}'\mathbf{\Sigma_{ij}^{-1}}\mathbf{y}\right]$$

The two-locus weighted conditional likelihood for each sibling pair is is therefore of

the form

$$L = \frac{\sum_{i=0}^{2}\sum_{j=0}^{2}[\mathbf{Q}]_{ij}L_{ij}}{\sum_{i=0}^{2}\sum_{j=0}^{2}[\mathbf{P}]_{ij}L_{ij}}$$

and the log-likelihood can be summed over all sibling pairs.

The implementation of this approach was fraught with numerical difficulties in optimisation, however. In particular, preliminary work suggested that this approach was extremely sensitive to starting values when the full 8 parameter model was specified, so much so as to preclude the method's usefulness. This approach was abandoned in favour of the hopefully more robust two-locus regression-based method, described in the next section.

8.1.2 An extended two-locus Haseman-Elston linkage method

This section considers an extension to the two-locus case of the recently reformulated Haseman-Elston regression method (Sham et al., 2002b; Sham and Purcell, 2001). In comparison to the original method (Haseman and Elston, 1972), the reformulation features two important changes: 1) rather than simply the squared difference between siblings, the trait variable is now a weighted composite of squared-sums and squared-differences between siblings, which can be shown to be approximately equivalent to variance components in power, 2) the regression equation is reversed to become the regression of IBD on trait, to ensure that the test is robust in selected samples. Although other authors have demonstrated multi-locus extensions to the original Haseman-Elston method (e.g. Tiwari and Elston, 1997b, reviewed in Chapter 5), this avenue has not been explored within the context of the reformulated Haseman-Elston.

The regression involves a composite of squared-sums and squared-differences weighted by the population sibling correlation, and eight IBD sharing variables, corresponding to the eight components of genetic variance for two QTL: additive effects (π_A and

π_B), dominance effects (z_A and z_B), additive × additive epistasis ($\pi_A\pi_B$), additive × dominance epistasis ($\pi_A z_B$ and $z_A\pi_B$) and dominance × dominance epistasis ($z_A z_B$).

Trait and IBD variables are standardised to allow the regression coefficients to be directly interpreted as variance components and to enable the analysis of selected samples. Each pair's trait score (the weighted sum of squared-sum and squared-difference) is adjusted according to the theoretical mean and variance, which depends upon the sibling correlation. The sibling correlation can either be estimated from the sample, or fixed to its population value if known. Similarly, for each sibling pair, each IBD sharing variable is mean-centred and standardised. The population mean and variance of the IBD sharing variables are given by basic genetic theory. The standardisation procedures are more involved than normal standardisation: the aim is to ensure that the regression coefficients directly estimate the components of variance. The details of these procedures are given below.

Method

The basic model is a multivariate regression of two-locus IBD sharing variables on a composite measure of squared trait difference and squared trait sum for siblings. When IBD sharing variables are the predictor variables, the substantial covariation between them is implicitly taken into account, such that their regression coefficients will represent the components of variance associated with the eight sources of genetic variance. However, in the present case, the multivariate regression of IBD sharing variables on the composite trait score does not yield regression coefficients that correspond to variance components, as this covariation is not addressed. Even if there were only additive effects at the first locus, nonzero regression coefficients would still be expected for the z_1, $\pi_1\pi_2$, $\pi_1 z_2$, $z_1\pi_2$ and $z_1 z_2$ variables as well as π_1, for these five variables are all strongly correlated with π_1. However, a transformation creates 8 new variables that are linear combinations of the 8 standardised IBD sharing vari-

ables: the regression coefficients for these new variables directly estimate the variance components. The standardisation and transformation procedures are described below.

Typically, the covariance matrix of the standardised and transformed IBD sharing variables, which is used in calculating the solution, would be estimated from the sample. In selected samples, this may produce biased results, however. To allow for the analysis of selected samples, this covariance matrix is fixed to the population values based on basic genetic theory, outlined in Chapter 5. The calculation of this matrix (which also depends on the sibling trait correlation) is detailed below. As the trait is the predictor variable, the procedure should be robust in selected samples.

Creating the composite score

The rationale behind the composite measure and its optimal performance is described in Sham and Purcell (2001). Basically, for a sibling pair with standardised trait scores T_1 and T_2, the composite trait index is defined

$$X = \frac{(T_1 + T_2)^2}{(1+r)^2} - \frac{(T_1 - T_2)^2}{(1-r)^2}.$$

Assuming that the trait is multivariate normal, it can be shown that the expected value of the composite is

$$E(X) = \frac{-4r}{1-r^2}$$

whilst its variance is

$$Var(X) = \frac{16(1+r^2)}{(1-r^2)^2}$$

and the covariance with the IBD sharing variable $\hat{\pi}$ is

$$\frac{4Q(1+r^2)Var(\hat{\pi})}{(1-r^2)^2}$$

as derived in Appendix C of Sham and Purcell (2001). As $\beta_{X\hat{\pi}} = Cov(X, \hat{\pi})/Var(\hat{\pi})$, then the regression coefficient will equal

$$\frac{4(1+r^2)}{(1-r^2)^2}Q$$

Dividing X by the first part of this factor will ensure that the regression coefficient is exactly Q, therefore the composite is 'standardised' X' as follows:

$$X' = [X - E(X)] \,/\, \left[\frac{4(1+r^2)}{(1-r^2)^2}\right]$$

The matrix $\mathbf{X}$ is then defined

$$[\mathbf{X}]_{ij} = \begin{cases} X' & i = j \\ 0 & i \neq j. \end{cases}$$

Standardisation of IBD sharing variables

As well as standardising the composite trait index, it is also necessary to standardise the IBD sharing variables. For the two-locus case allowing for all orders of epistatic interaction, the vector

$$\hat{\mathbf{\Pi}} = \left[\hat{\pi}_1 \quad \hat{z}_1 \quad \hat{\pi}_2 \quad \hat{z}_2 \quad \hat{\pi}_1\hat{\pi}_2 \quad \hat{\pi}_1\hat{z}_2 \quad \hat{z}_1\hat{\pi}_2 \quad \hat{z}_1\hat{z}_2\right]$$

represents the 8 allele-sharing variables, the mean and variance of which are

$$\mathbf{E}_{\hat{\mathbf{\Pi}}} = \left[\frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{2} \quad \frac{1}{4} \quad \frac{1}{4} \quad \frac{1}{8} \quad \frac{1}{8} \quad \frac{1}{16}\right]$$

and

$$\mathbf{V}_{\hat{\mathbf{\Pi}}} = \left[\frac{1}{8} \quad \frac{3}{16} \quad \frac{1}{8} \quad \frac{3}{16} \quad \frac{5}{64} \quad \frac{5}{64} \quad \frac{5}{64} \quad \frac{15}{256}\right]$$

respectively. Each element is then standardised by the appropriate mean and variance to give the vector $\hat{\mathbf{\Pi}}_\mathbf{C}$ of standardised allele-sharing variables.

Adjustment for putting allele-sharing variables as the dependent

As the regression coefficient of X' on $[\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i}$ is

$$\beta_{X'[\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i}} = \frac{Cov([\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i}, X')}{Var([\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i})} = Q$$

it follows that the regression coefficient of $[\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i}$ on X' is

$$\beta_{[\hat{\mathbf{\Pi}}_\mathbf{C}]_\mathbf{i}X'} = Q\frac{1}{Var(X')}$$

and so it is necessary to make a further adjustment to the IBD variables, $\hat{\mathbf{\Pi}}_\mathbf{C}$, multiplying each element by the variance of the composite index

$$\begin{aligned} Var(X') &= Var\left(\frac{X}{\left[\frac{4(1+r^2)}{(1-r^2)^2}\right]}\right) = \frac{1}{\left[\frac{4(1+r^2)}{(1-r^2)^2}\right]^2}\left[\frac{16(1+r^2)}{(1-r^2)^2}\right] \\ &= \frac{1}{\left[\frac{(1+r^2)}{(1-r^2)^2}\right]} \end{aligned}$$

in order to account for the reversing dependent and independent variables. The combined standardisation can be represented by the matrix operations

$$\hat{\mathbf{\Pi}}_\mathbf{C} = \mathbf{S}^{-1}(\hat{\mathbf{\Pi}} - \mathbf{E}_{\hat{\mathbf{\Pi}}})$$

where $\mathbf{S}$ is a diagonal 8×8 matrix for which

$$[\mathbf{S}]_{ii} = \frac{[\mathbf{V}_{\hat{\mathbf{\Pi}}}]_i}{Var(X')}.$$

Transforming the IBD variables

As mentioned above, the regression of $\mathbf{\Pi_C}$ on $\mathbf{X}$ would fail to account for the substantial covariation between IBD sharing variables. A transformation is proposed which creates 8 linear combinations of the IBD sharing variables, which estimate the eight components of variance in a two-locus epistatic scenario directly, i.e. additive and dominance variance components for each locus, and four epistatic variance components representing additive × additive, additive × dominance, dominance × additive and dominance × dominance interaction.

Let $\mathbf{\Sigma_{\hat{\Pi}}}$ be the covariance matrix for the 'standardised' IBD variables. Then the transformation matrix $\mathbf{T}$ is defined

$$\mathbf{T} = \mathbf{\Sigma_{\hat{\Pi}}^{-1}} \mathbf{D}$$

where $\mathbf{D}$ is an 8×8 diagonal matrix containing the variances of the IBD variables. In this case, the vector $\mathbf{Y}$ of transformed IBD variables is

$$\mathbf{Y} = \mathbf{T}\hat{\mathbf{\Pi}}_\mathbf{C}$$

The rationale for this transformation is as follows. If $[\mathbf{T}]_i$ is the i^{th} row of the transformation matrix $\mathbf{T}$, then it is required that the regression of

$$[\mathbf{T}]_i' \mathbf{S^{-1}} (\hat{\mathbf{\Pi}} - \mathbf{E_{\hat{\Pi}}})$$

on $\mathbf{X}$ to have regression coefficient Q_i, i.e. to directly estimate the component of variance. That is,

$$\begin{aligned} Q_i &= \frac{Cov\left([\mathbf{T}]_i' \mathbf{S^{-1}} (\hat{\mathbf{\Pi}} - \mathbf{E_{\hat{\Pi}}}), X'\right)}{Var(X')} \\ &= [\mathbf{T}]_i' \mathbf{D^{-1}} \mathbf{\Sigma_{\hat{\Pi} X'}} \end{aligned}$$

In matrix terms, the vector of regression coefficients β which directly represent the variance components is

$$\beta = \mathbf{T}\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}\mathbf{X}'}.$$

As $\beta = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}\mathbf{X}'}$ then

$$\beta = \mathbf{T}\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}\beta$$

and therefore

$$\mathbf{T}\mathbf{D}^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}} = \mathbf{I}$$

where $\mathbf{I}$ is an 8×8 identity matrix. Rearranging gives $\mathbf{T} = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\Pi}}}^{-1}\mathbf{D}$.

Fixing the IBD covariance matrix to population value

The covariance matrix of the 8 original IBD variables, $\boldsymbol{\Sigma}_{\boldsymbol{\Pi}}$, is given by basic genetic theory as described in Chapter 5. The covariance matrix of the standardised and transformed IBD variables is calculated

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathbf{T}\mathbf{S}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\Pi}}\mathbf{S}^{-1}\mathbf{T}'$$

where $\mathbf{S}$ and $\mathbf{T}$ are defined above.

Solving the regression equation

When $\boldsymbol{\Sigma}_{\mathbf{Y}}$ is known, as in the present case, the solution to the regression equation is obtained by summing over all N sibling pairs (Crowder and Hand, 1990)

$$\hat{\beta} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{Y}_i\right)$$

with covariance matrix

$$\mathbf{V}_{\beta} = \left(\sum_{i=1}^{N}\mathbf{X}_i'\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathbf{X}_i\right)^{-1}$$

Reduced model

As well as the full two-locus model, containing all four epistatic terms, a reduced 5-parameter model was also fitted to the data. The reduced model contains 4 main effect terms, but only a single epistatic parameter, representing additive × additive epistatic interaction. The matrices **S**, **T** and therefore $\mathbf{\Sigma_Y}$ are all recalculated on the basis of only these 5 terms. The reduced model allows a 1 degree of freedom test of any epistatic effect, as opposed to a 4 degree of freedom test for all epistatic effects. As will be seen, this formulation is preferable in most circumstances, although the model misspecification (i.e. if the higher-order epistatic terms contribute to the variance strongly) can lead to biased parameter estimates.

Testing linear hypotheses

For the linear hypothesis $\mathbf{H}\beta = \mathbf{h}$ the test statistic is

$$T_H = (\mathbf{H}\hat{\beta} - \mathbf{h})'(\mathbf{H}\mathbf{V}_\beta\mathbf{H}')^{-1}(\mathbf{H}\hat{\beta} - \mathbf{h})$$

which has an approximate χ^2 distribution in large samples (Crowder and Hand, 1990). If there are p terms in the full model and $p - r$ terms in the restricted model, then **H** is a $r \times q$ matrix and **h** is a $r \times 1$ vector. All elements of **h** are set equal to zero (i.e. to test the hypothesis that the restricted terms equal 0). If the first restricted term is i in the full model is restricted in the submodel, then the element $(1, i)$ is set to 1, etc. For example, to test the linear hypothesis that there are no epistatic effects above additive × additive interaction

$$\mathbf{H} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and $\mathbf{h} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}'$. This test statistic is distributed as χ^2_3.

8.2 Simulations

8.2.1 Overview

Samples of sibling pairs were simulated under a variety of epistatic models. In all cases, a bivariate normal trait was simulated for the sibling pair, along with perfect information IBD values for two unlinked loci. Epistatic models were specified in terms of a matrix of 9 genotypic means and two allele frequencies, as described in Chapter 5. In all cases an unselected sample size of 5000 sibling pairs and a selected sample size of 500 sibling pairs was used, i.e. 10% selection. Samples were selected using the method described in Chapter 2. In all cases, the two loci jointly accounted for 40% of the trait variance. The residual sibling correlation was varied, by either simulating residual shared and nonshared effects to account for either 20% and 40% of the total variance respectively, or 40% and 20%. For each model, the expected sibling correlation was calculated, as described in Chapter 5 – in analysis, the sibling correlation was fixed to this value. Each condition was replicated 2000 times.

Two different models were applied to the data: a full model with 8 parameters (i.e. including the higher order epistatic terms featuring dominance) and a reduced 5 parameter model (including additive and dominance main effects but only additive × additive epistasis).

For the full model, three tests were conducted: 1) an 8 degree of freedom test of no QTL effects (all parameters equal zero) 2) a 4 degree of freedom test of no epistatic effects (the four regression coefficients corresponding to the four epistatic variance components equated to zero) and 3) a 3 degree of freedom test of no higher-order epistatic effects (the three regression coefficients corresponding to the three higher-order epistatic variance components). For the reduced 5 parameter model, two tests

were conducted: 1) a 5 degree of freedom test of no QTL effects and 2) a 1 degree of freedom test of no epistatic effect.

The four different epistatic models investigated in these analyses were:

$$M1 = \begin{bmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \\ 0 & 0.5 & 1 \end{bmatrix} \quad M4 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad M8 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad M12 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0.5 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

although model $M4$ was simulated using three different allele frequencies (0.1, 0.5 and 0.9) for the first alleles (i.e. making the double-homozygote (with the mean of 1) either quite common, fairly rare or very rare, respectively, labelled $M4_A$, $M4_B$ and $M4_C$). That is, the frequency of the double increaser homozygote would be around 65%, 6% or 0.1%. In this last extreme case, although the QTL are simulated so as to account for a constant proportion of variance (and therefore, the actual effects would be very large when the frequency is so low), one would expect this scenario to highlight the difference between theoretical expectation and the analysis of finite samples.

8.2.2 Results

The regression analyses were also performed on the untransformed IBD variables as well as the transformed ones: as expected, the untransformed analyses failed to distinguish between epistatic and non-epistatic effects. In contrast, Tables 8.1 and 8.2 illustrate the results for the transformed analysis (displaying the average test statistics and power at $\alpha = 0.05$ respectively), which clearly differentiates between epistatic and non-epistatic effects. The tests of any QTL effect ('All') are largely all highly significant, which is to be expected as the QTL were simulated to account for almost half the trait variance. The average test statistic for any QTL effect varies considerably from model to model, however. This is largely due to differences in the

sibling correlation implied by the different models, which is known to influence the power of the linkage test. However, model $M4_C$ shows a distinctly different pattern, with the test statistics at their expected values under the null. This is undoubtedly because the double homozygote would be very unlikely to occur even in samples of 5000 sibling pairs, because it is so rare. Therefore, in the majority of the samples, there would not be any variation in the trait due to either QTL. This extreme example perhaps just makes the more general point, that considering the expected variance components alone is not enough to accurately predict performance (although with a much larger number of replicates, results should converge to their expected values). A similar, but less extreme, trend can also be seen between models $M4_A$ and $M4_B$, where $M4_A$ is more powerful as the double increaser homozygote will occur more frequently in this scenario.

Comparing the test statistics in full and selected samples, in most cases approximately 60% of the information for linkage is retained by 10% of the sibling sample. Model $M4_B$ shows a distinctly different pattern, however, with all the information being retained in the selected sample. In general, selection is likely to be more efficient when the same amount of variation is characterised in terms of rare but large effects as opposed to small but common effects. This is the case in model $M4_B$ – all of the small number of sibships containing at least one individual with the double increaser homozygote genotype are going to be informative, whereas all other sibships will not be. This is represented in the lower full sample test statistic also. Epistasis, in and of itself, does not guarantee any particular pattern with respect to this effect size / effect frequency continuum. Model $M12$, the 'most epistatic' of the models considered here does not show this pattern, and consequently selection is less efficient.

Looking at the results for the specific tests of epistasis in both Tables reveals an equally ambiguous pattern of results. Additive model $M1$ does not have any epistatic QTL effects, and, as established above, model $M4_C$ barely has *any* QTL effects, so

	Full 8 parameter model						Reduced 5 parameter model			
	All (8 df)		Epi (4 df)		Epi (3 df)		All (5 df)		Epi (1 df)	
	F	S	F	S	F	S	F	S	F	S
$M1$	166.77	107.16	4.14	4.31	3.05	3.18	163.72	103.98	1.09	1.13
$M1$	316.22	193.50	4.31	4.40	3.09	3.16	313.12	190.34	1.21	1.24
$M4_A$	79.27	47.58	4.06	4.06	2.95	3.02	76.32	44.56	1.11	1.05
$M4_A$	138.52	76.15	4.46	4.36	3.04	3.08	135.49	73.08	1.42	1.29
$M4_B$	50.47	51.27	8.89	8.86	3.78	3.76	46.69	47.51	5.11	5.10
$M4_B$	57.11	55.37	8.95	8.69	3.37	3.32	53.74	52.05	5.58	5.36
$M4_C$	7.83	7.78	3.88	4.05	2.94	3.02	4.89	4.76	0.94	1.03
$M4_C$	7.98	8.06	3.97	4.04	3.09	3.07	4.89	4.99	0.89	0.97
$M8$	78.35	42.23	4.19	4.05	2.95	2.98	75.41	39.24	1.25	1.06
$M8$	134.80	63.82	4.35	3.79	2.88	2.69	131.92	61.13	1.47	1.10
$M12$	61.40	36.83	23.27	14.69	7.45	5.20	53.95	31.63	15.83	9.49
$M12$	90.60	42.57	33.03	18.37	10.15	6.04	80.45	36.54	22.88	12.33

Table 8.1: Transformed test statistics for two-locus Haseman-Elston method: for full and selected samples (F and S). Top of each pair of rows: lower sibling correlations; bottom of each pair: higher sibling correlation. See text for further details.

performance of the epistasis tests is near the expectation under the null in both cases. Arguably the tests are a little anti-conservative – the Type I error rates for $M1$ which should have no epistasis are a little high. Otherwise, whether or not a given model will show detectable epistatic effects or not is largely due to the associated components of variance, as established in Chapter 5.

Figure 8.1 illustrates the expected variance components under the true model (the left column, with each row corresponding to models $M1$ – $M12$ going from top to bottom). Results are given for both full and selected samples, under high and low residual sibling correlations (H and L). As can be seen, models $M4_A$ and $M8$ are actually largely additive in terms of the magnitudes of the components of variance. In contrast, $M4_B$ and $M12$ show much stronger epistatic variance components. As shown in the Tables, epistasis is detectable in these latter two conditions, but not in the first two. For the factors mentioned in Chapter 5, power to detect these effects is still relatively low, however.

The middle and right columns of Figure 8.1 represent the average estimated variance components under the full 8 parameter and reduced 5 parameter model respectively. With the exception of $M4_C$, the variance components are unbiased estimate in

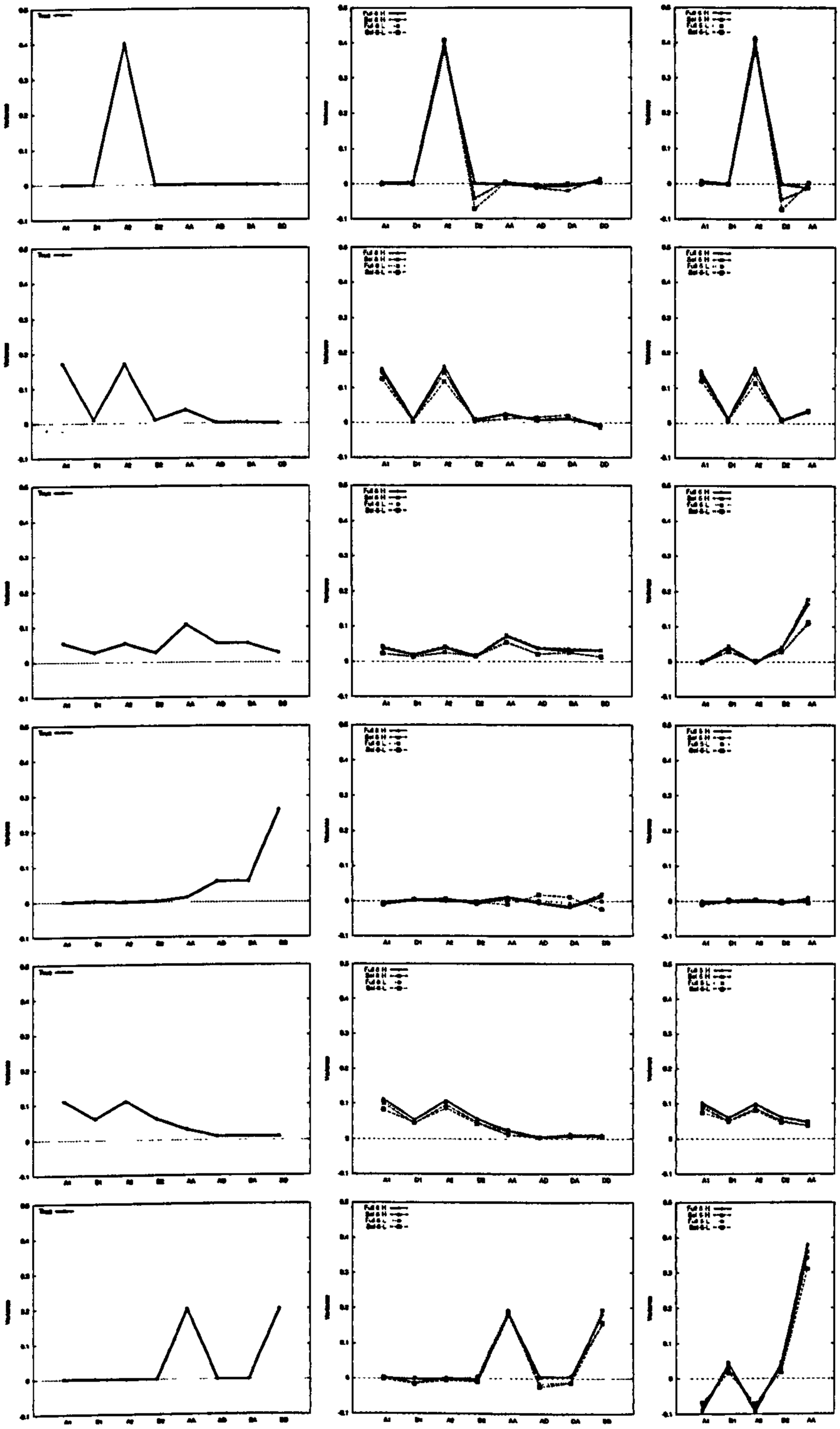


Figure 8.1: Average estimated variance components from sibling pair linkage. Rows from top to bottom represent models $M1$, $M4_A$, $M4_B$, $M4_C$, $M8$ and $M12$. The left column represents the expected variance components; the middle column represents the components estimated under the full model; the right column represents the components under the reduced model. In each plot, the four lines represent the four combinations of full and selected samples / high and low sibling correlation.

	Full 8 parameter model						Reduced 5 parameter model			
	All (8 df)		Epi (4 df)		Epi (3 df)		All (5 df)		Epi (1 df)	
	F	S	F	S	F	S	F	S	F	S
$M1$	1.00	1.00	0.06	0.08	0.05	0.07	1.00	1.00	0.06	0.07
$M1$	1.00	1.00	0.09	0.09	0.07	0.07	1.00	1.00	0.07	0.08
$M4_A$	1.00	1.00	0.05	0.06	0.05	0.06	1.00	1.00	0.07	0.06
$M4_A$	1.00	1.00	0.07	0.07	0.05	0.06	1.00	1.00	0.10	0.09
$M4_B$	1.00	1.00	0.39	0.39	0.09	0.09	1.00	1.00	0.53	0.52
$M4_B$	1.00	1.00	0.39	0.36	0.07	0.08	1.00	1.00	0.58	0.55
$M4_C$	0.03	0.05	0.03	0.05	0.03	0.06	0.02	0.05	0.10	0.04
$M4_C$	0.03	0.05	0.02	0.05	0.03	0.05	0.02	0.05	0.12	0.05
$M8$	1.00	1.00	0.07	0.05	0.05	0.06	1.00	1.00	0.09	0.06
$M8$	1.00	1.00	0.06	0.04	0.04	0.04	1.00	1.00	0.10	0.06
$M12$	1.00	1.00	0.96	0.77	0.41	0.20	1.00	1.00	0.97	0.81
$M12$	1.00	1.00	1.00	0.93	0.63	0.27	1.00	1.00	0.99	0.92

Table 8.2: Power for transformed two-locus Haseman-Elston method: for full and selected samples (F and S). Top of each pair of rows: lower sibling correlations; bottom of each pair: higher sibling correlation. See text for further details.

all cases under the 8 parameter model. As explored in Chapter 5, the reduced 5 parameter model (representing, in some cases, a misspecification of the true model) can lead to parameter estimates that are biased (e.g. the negative estimates of additive variance components for $M12$). Another point of interest is that for the additive-only $M1$ model, sample selection tends to bias the estimate of the dominance variance component to be negative.

Finally, note that despite the very high epistatic variance associated with model $M4_C$, the estimated variance components are all near zero, for reasons mentioned above. In general, for this level of selection, it appears that the properties of selected samples with respect to the detection of epistasis are very similar to those of unselected samples.

8.3 Epistasis in QTL association analysis

As shown in Chapter 7, it is possible to incorporate a moderator variable into the variance components association model described in Chapter 3. In this section, a second locus rather than an environmental measure is introduced as a moderator of

the QTL effect. Such a procedure allows for tests of gene × gene interaction in selected sibship samples.

To perform a full analysis of two loci within the conditional framework, it would be necessary to consider all possible genotypic configurations, GC, as two-locus sibship genotypes and inheritance vectors. That is, for two loci G_1 and G_2,

$$L(G_1, G_2|X) = \frac{L(X|G_1, G_2)L(G_1, G_2)}{\sum_{G_1}\sum_{G_2} L(X|G_1, G_2)L(G_1, G_2)}.$$

Considering every possible two-locus genotypic configuration in the denominator may well become computationally prohibitive, however, especially with larger sibships. An alternative strategy, used here, is to regard the second locus as a modifier variable, in the same way as $G \times E$ analysis. The likelihood is therefore

$$L(G_1|X, G_2) = \frac{L(X|G_1, G_2)L(G_1|G_2)}{\sum_{G_1} L(X|G_1, G_2)L(G_1|G_2)}$$

which, assuming G_1 and G_2 are in linkage equilibrium and unlinked, equals

$$L(G_1|X, G_2) = \frac{L(X|G_1, G_2)L(G_1)}{\sum_{G_1} L(X|G_1, G_2)L(G_1)}.$$

If the loci were linked or in linkage disequilibrium, $L(G_1|G_2)$ would have to be properly specified – this section restricts attention to the simpler case when the two loci are unlinked and in linkage equilibrium.

This analysis is essentially of a single locus, the effects of which might be modified by alleles at a second locus. As a consequence, analogous to the inability to estimate the main effect of a covariate in the conditional approach, any main effects of the modifier locus will not be detected. Given this 'asymmetry' (i.e. one locus must be chosen as the dependent, one as the modifier), the locus with the strongest main effect in standard analysis should be made the dependent test locus.

Assuming the second locus is diallelic, two dummy moderator variables are created, A_M and D_M, coded 1, 0 and -1 and 0, 1, 0 for additive and dominance effects respectively, according to genotype. These variables must be standardised using the population mean and variance, i.e. the population allele frequency must be known. The coding is then represented as A_1, A_0 and A_{-1}, and D_0, D_1, D_0. The sibling correlation for A_M and D_M must also be specified: these are simply 0.5 and 0.25 for full sibling pairs.

The parameters for the first covariate, A_M, are α_M, α_A and α_D. The parameters for the second covariate, D_M, are δ_M, δ_A and δ_D. For example, α_A represents an additive $\times$ additive interaction effect; δ_A represents the dominance component of the modifier locus interacting with the additive effect of the test locus – an additive effects at the test locus $\times$ dominance effects at the modifier locus interaction. These six parameters, along with the main effects at the test locus a and d, combine to form the eight effects listed in Chapter 5. Reconstructing the 3 $\times$ 3 table of genotypic means has to take account of the standardisation of A_M and D_M, however. For test locus A and modifier locus M, the 9 genotypic means are

$$
\begin{aligned}
\mu_{AA/MM} &= a + \alpha_M A_1 + \delta_M D_0 + \alpha_A A_1 + \delta_A D_0 \\
\mu_{AA/Mm} &= a + \alpha_M A_0 + \delta_M D_1 + \alpha_A A_0 + \delta_A D_1 \\
\mu_{AA/mm} &= a + \alpha_M A_{-1} + \delta_M D_0 + \alpha_A A_{-1} + \delta_A D_0 \\
\mu_{Aa/MM} &= d + \alpha_M A_1 + \delta_M D_0 + \alpha_D A_1 + \delta_D D_0 \\
\mu_{Aa/Mm} &= d + \alpha_M A_0 + \delta_M D_1 + \alpha_D A_0 + \delta_D D_1 \\
\mu_{Aa/mm} &= d + \alpha_M A_{-1} + \delta_M D_0 + \alpha_D A_{-1} + \delta_D D_0 \\
\mu_{aa/MM} &= -a + \alpha_M A_1 + \delta_M D_0 - \alpha_A A_1 - \delta_A D_0 \\
\mu_{aa/Mm} &= -a + \alpha_M A_0 + \delta_M D_1 - \alpha_A A_0 - \delta_A D_1 \\
\mu_{aa/mm} &= -a + \alpha_M A_{-1} + \delta_M D_0 - \alpha_A A_{-1} - \delta_A D_0.
\end{aligned}
$$

8.3.1 Simulations

The above method, implemented in the `cafe` program, is applied to eight two-locus models, as illustrated in Figure 8.2. The first four models (reading across columns, then down rows) are from the previous epistatic linkage simulations: models M_1, M_4, M_8 and M_{12}. In all cases, both loci are diallelic with equifrequent alleles. The remaining four epistatic models represent specific orders of epistasis, against the background of an additive main effect at the test locus: that is, additive × additive, additive × dominance, dominance × additive and dominance × dominance.

The full sample size is 1000 individuals, or 500 sibling pairs; selected samples contain 200 individuals, or 100 sibling pairs. The two loci account for 30% of the trait variance in all cases. For each condition, the results represent only a single replicate – more extensive simulations are planned for the future.

Table 8.3 presents the likelihood ratio test statistic for each condition. Although it is unwise to generalise from single replicates, the results look promising. For a 4 degree of freedom test, the LRT are small under model 1, for which there are no epistatic effects, but much higher under all other models, for which there are epistatic effects. There appears to be no consistent differences between conditional and unconditional approaches in terms of the LRTs; singletons appear to have larger LRTs than pairs.

The main purpose of presenting these preliminary results is to demonstrate some problems with the model, however. There is no result for the model 4 conditional test with selected singletons. This is because it so happened that not all 9 genotypes were represented in this selected sample. Just as a model including a dominance effect would not be identified if only the two homozygotes were observed in a sample, the full epistatic model is not identified in this case. This problem is likely to arise often in selected samples, especially when allele frequencies are rare. Data should be carefully inspected before fitting the full epistatic model in all cases.

Figures 8.3 and 8.4 show the estimated 9 genotype means for the two-locus model,

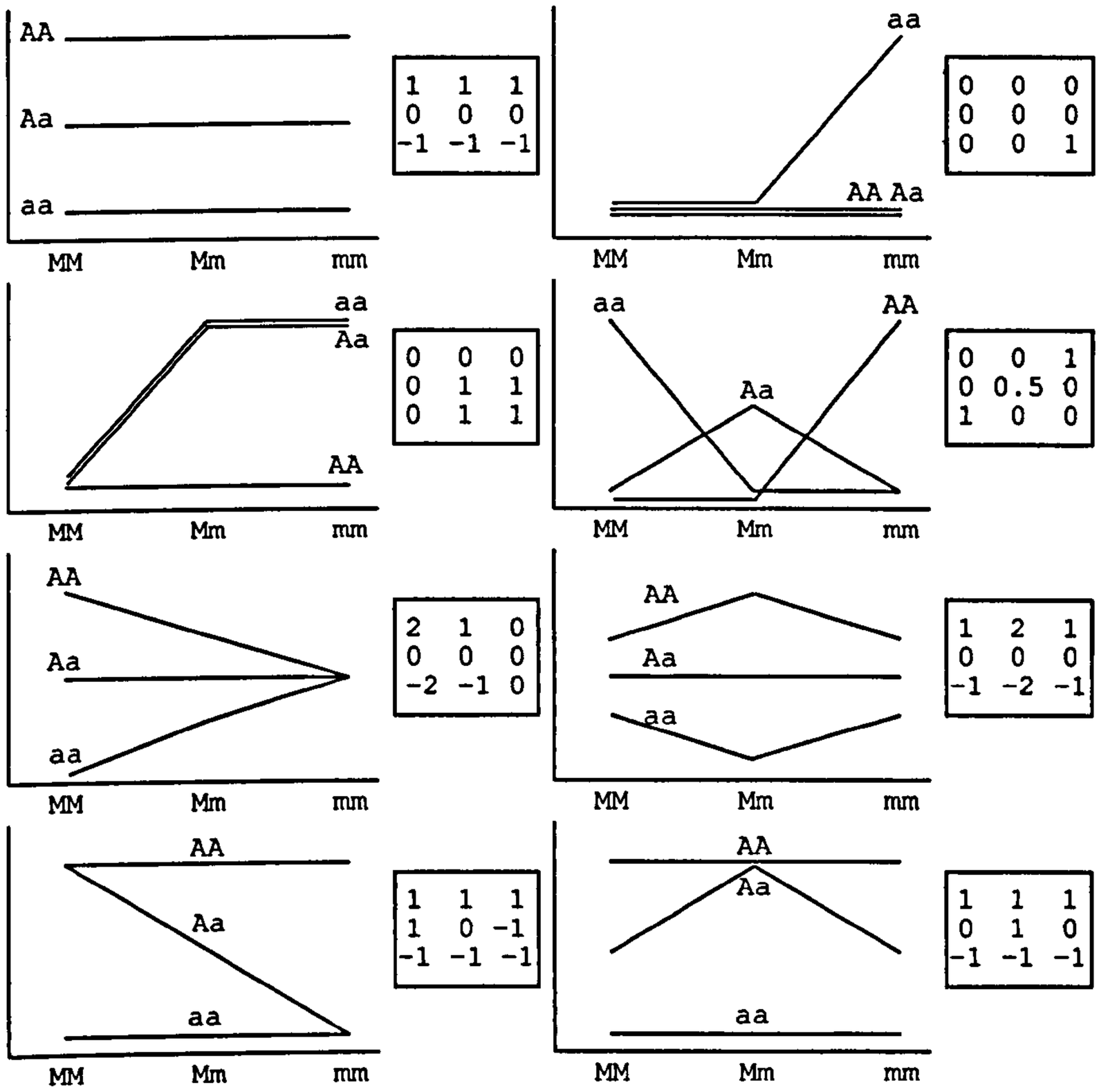


Figure 8.2: Schematic views of the 8 epistatic models used: see text for details.

	Full sample				Selected sample			
	$s = 1$		$s = 2$		$s = 1$		$s = 2$	
Model	NC	C	NC	C	NC	C	NC	C
1	2.51	2.86	5.88	5.33	3.24	6.53	4.77	3.95
2	118.15	120.57	81.47	73.96	73.29	93.03	46.12	45.67
3	49.68	52.25	28.43	27.17	71.17	46.45	28.41	17.43
4	333.68	327.02	181.12	175.24	228.47		124.37	117.92
5	78.30	79.53	44.20	44.63	71.47	78.49	29.96	26.49
6	21.90	22.40	16.02	16.06	24.47	21.42	12.42	10.25
7	70.96	68.75	35.33	32.91	62.82	46.50	35.69	26.89
8	31.96	32.53	29.62	26.45	25.03	22.54	26.05	20.08

Table 8.3: Test statistics (4 degree of freedom test of epistasis) for two-locus association model: for full and selected samples, singletons ($s = 1$) and sibling pairs ($s = 2$).

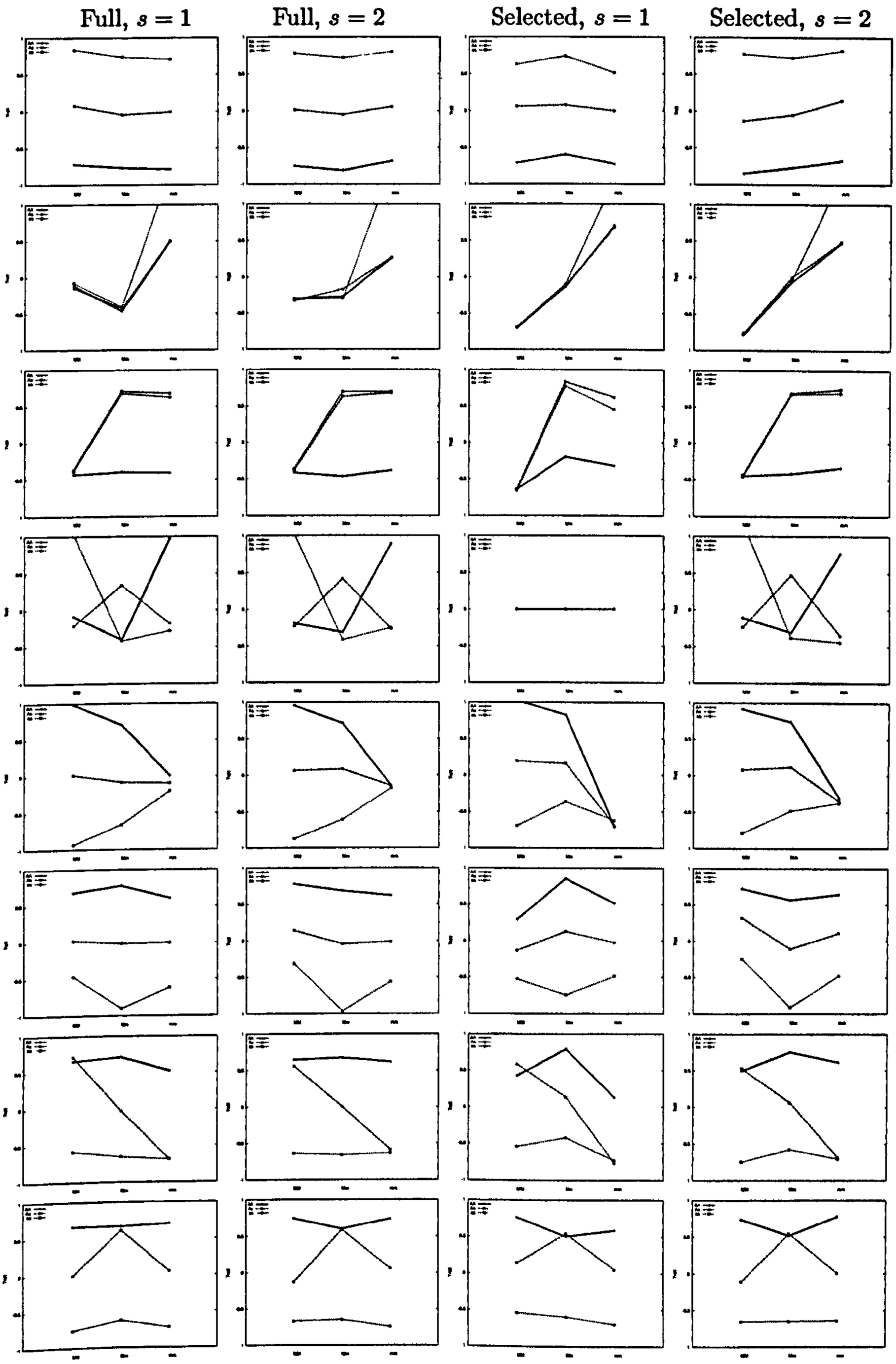


Figure 8.3: Estimated genotypic means from association analysis: conditioning. Each row represents a model as described in Figure 8.2.

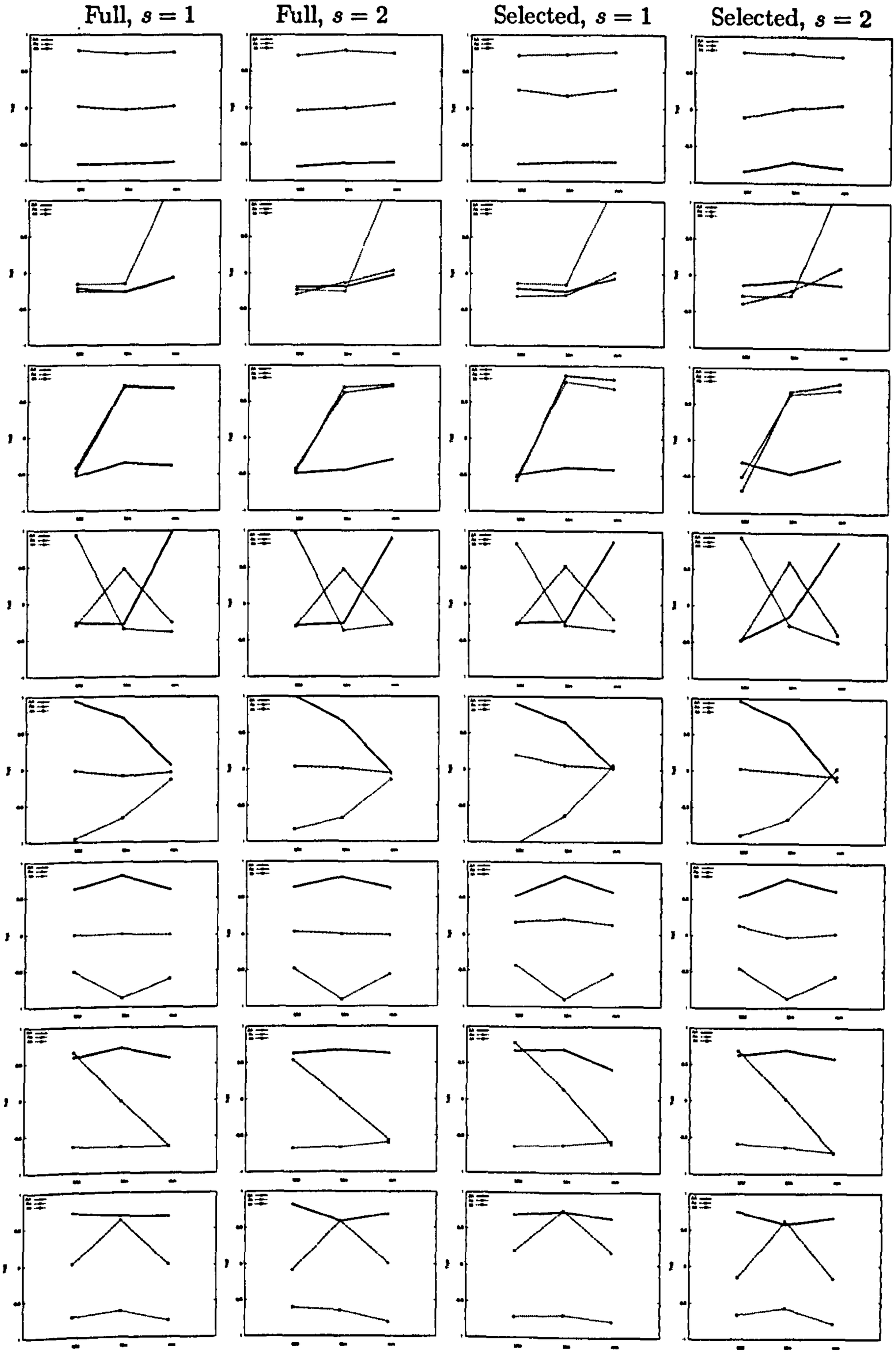


Figure 8.4: Estimated genotypic means from association analysis: not conditioning. Each row represents a model as described in Figure 8.2.

based on the conditional and unconditional tests respectively. In general, both methods do a good job of recovering the two-locus effects. However, certain models (e.g. the second model) show signs of biased estimates in the unconditional approach. This is in fact a further problem of identification. Conditional on the moderator (i.e. the second locus), when there is no main effect of the first locus, then β_M and δ_M cannot be properly identified. As a consequence, just the *relative differences* in first locus genotypes are correct. This issue never arose in the $G \times E$ analyses presented in the previous Chapter because the moderator was continuous (even though at certain values of the moderator there was no main effect of the first locus). Although this should not bias the LRT values, it does mean that care must be taken in interpreting the expected two-locus means model from the conditional analysis. Further research is required in order to make this approach generally applicable.

8.4 Summary

This Chapter has presented two-locus linkage and association models that are robust in selected samples. For the scenarios simulated in this Chapter, the differences between full and selected samples have not been particularly marked. For linkage, this will probably hold for a wide range of models – as demonstrated in Chapter 5, additive and epistatic effects are, in any case, blurred in QTL linkage analysis. For association studies, epistatic interaction effects will probably be harder to detect in selected samples, especially when more extreme selection strategies are employed and allele frequencies are rarer. That is, although additive effects can be estimated from extreme-scoring high and low homozygotes alone, to detect epistasis it is necessary to sample the majority of the 9 two-locus genotypes – selecting on trait cannot esure this.

Chapter 9

Selection & population stratification

This Chapter considers the potential impact of population stratification on quantitative trait locus (QTL) association tests using phenotypically selected samples of unrelated individuals. As well as a standard regression-based test of QTL association, a novel maximum likelihood method is developed that should be robust in selected samples. Simulation results are used to evaluate both methods and to explore some novel problems arising when dealing with stratification in selected samples. Overall, both methods perform well: although the regression method is often marginally more powerful, the maximum likelihood method provides valid estimates of QTL effect size and is more robust in samples selected from non-normal distributions.

9.1 Background

As discussed in Chapter 6, it is possible to detect stratification within a sample by analysing a number of unlinked markers for signs of linkage disequilibrium, reflecting stratification effects. Furthermore, it is possible to assign individuals probabilistically to each of the K classes, or strata, considered in the best-fitting solution. The second

issue, how to correct for the detected population substructure, is explored in this Chapter. In particular, issues arising when using phenotypically selected samples are considered.

The assumed experimental design is as follows: *(i)* phenotype a large number of unrelated individuals at random, *(ii)* select phenotypically extreme individuals (e.g. top and bottom 10%), *(iii)* genotype selected individuals at a number of marker loci (which include both the test loci and the 'null' markers to feature in stratification analysis – these may overlap, partially or completely), *(iv)* perform stratification analysis, *(v)* perform association analysis, potentially conditional on stratum membership as determined in stage *(iv)*.

In many of the simulations below, stage *(iv)* is skipped and perfect knowledge of stratum membership is assumed. That is, rather than estimating posterior probabilities of class membership from unlinked marker data, the 'posterior probabilities' are actually binary 0/1 variables corresponding to whether an individual was simulated as belonging to that class or not. This strategy is adopted for a number of reasons, not least the practical reason of speeding up the simulations considerably. However, as shown in Chapter 6, if enough marker loci are employed, classification can be (near) perfect in any case. Furthermore, perfect classification represents a 'worst case scenario' so far as certain phenomena observed in selected samples are concerned, as will be illustrated below.

As well as the ML method, a standard regression-based approach is described. The implementation of the ML method in the computer package `L-ASSOC` is also covered.

The remainder of the Chapter reports the simulation studies used to evaluate both methods under a number of conditions, with particular emphasis on performance in selected samples when modelling population stratification effects. The various conditions investigated include: additive and dominance QTL effects; alternate selective sampling schemes; testing for mean differences between strata only; testing for QTL ×

strata interaction; testing for allele frequency differences between strata only; the impact of unequal class size; the impact of population outliers; the impact of correcting for the stratum-specific means; the impact of imperfect classification.

As mentioned, for many of the simulations outlined above, stage *(iv)* is skipped and the posterior probabilities are directly specified during sample simulation. The simulations looking at the impact of imperfect classification go some way to mimicking how an under-powered stratification analysis might influence results.

A final section considers the potential implications of performing the stage *(iv)* stratification analysis in a selected sample. In particular, it is shown that selecting the extremes of a polygenic trait before performing the stratification and association analyses can lead to a considerable reduction in power.

9.2 Testing for association in selected samples

Typically in tests of association, the trait is the dependent variable whilst the predictor variables are some function of genotype. A simple linear regression approach might model an association between a quantitative trait and a diallelic locus

$$T_i = \mu + \beta_A A_i + \beta_D D_i + \varepsilon_i$$

where individual i's trait score T_i is a function of additive effects A_i coded 1, 0 and -1 to represent alleles A_1A_1, A_1A_2 and A_2A_2; dominance effects are coded by D_i as 0, 1 and 0; μ is the intercept and ε_i is the residual error. Likewise, standard variance components models of association (e.g. Fulker et al., 1999) model the likelihood of observing the trait given the individual's genotype, $L(T|G)$.

However, such approaches are not necessarily valid for the analysis of phenotypically selected samples. In general, for any two variables X and Y, if Y is modelled as dependent on X, then it is acceptable to select a sample on X only, but not on Y.

That is, when Y is the dependent it is modelled conditional on X, i.e. $E(Y|X)$ in the regression framework, or $L(Y|X)$ in the variance components framework. Therefore, selecting on X will not basis the estimates, although it may lead to a reduction in power.

Genetic studies will typically select samples on the basis of individuals' trait scores, not their genotypes. In order to obtain valid estimates of the QTL variance, a test of association in selected samples should therefore have genotype as its dependent variable. For example,

$$A_i = \mu + \beta_T T_i + \varepsilon_i$$

However, within the straightforward regression framework, it is not clear how to include both additive and dominance effects on the left hand side of the equation, and the residuals will not be normally distributed. For this reason, an alternative maximum likelihood approach, such as the one described below, may be preferable.

9.2.1 Maximum likelihood model

A full likelihood model of both trait and genotype involves their joint probability, $P(T \bigcup G)$, which can be re-expressed as either $P(T|G)P(G)$ or $P(G|T)P(T)$. As mentioned above, for selected samples it is preferable to avoid the first formulation, which involves modelling the trait conditional on genotype, $P(T|G)$. The current approach therefore evaluates the likelihood of observing an individual's genotype conditional on trait score – the second formulation $P(G|T)P(T)$. This 'conditioning-on-trait values' approach has been previously adopted in the context of complex segregation analysis (Ewens and Shute, 1986) and variance component linkage (Sham et al., 2000a). The $P(T)$ component can be ignored when formulating the likelihood – as the data have been selected, $P(T)$ will also reflect the ascertainment process, which it might not be possible to model easily, and, in any case, it would cancel in any likelihood ratio test.

To allow for stratification effects, association is modelled conditional on belonging to class j of K discrete classes. For each individual, the probabilities of belonging to each class will be the posterior probabilities produced by a program such as L-POP from genetic background information. Alternatively, these 'probabilities' could be binary variables coded 0/1 based on some other classification scheme, such as self-reported ethnicity. The posterior probabilities are denoted $P(C|G)$. The class-conditional likelihood will be based on $P(G|T,C)$. The overall likelihood will be the weighted sum $\sum_j P(G|T,C_j)P(C_j|G)$ therefore.

A QTL may differ in allele frequency between classes; additionally, a QTL may have different effects in different classes, i.e. a QTL × class interaction. These possibilities are incorporated in the model described below. It is also possible for there to be differences in trait means between classes – indeed, there has to be both trait differences and allele frequency differences for stratification to occur. However, within the current framework, it is not possible to specifically estimate such class-specific effects on the trait, as these are confounded with QTL frequency and effect. That is, for two classes, a mean class difference with no QTL effect could produce the same pattern of results as a QTL with an additive effect and an allele frequency of 0 in one class, 1 in the other class. Class means can be estimated from the data or calculated from the other class-specific parameters in the model.

The model is parameterised in terms of class-specific additive genetic values (a_j), dominance deviations (d_j) and allele frequencies (p_j). Mean-centred class-specific genotypic means are calculated

$$\begin{aligned}
\mu_{11|j} &= a_j - (a_j(p_j - q_j) + 2p_jq_jd_j) \\
\mu_{12|j} &= d_j - (a_j(p_j - q_j) + 2p_jq_jd_j) \\
\mu_{22|j} &= -a_j - (a_j(p_j - q_j) + 2p_jq_jd_j)
\end{aligned}$$

and class-specific genotype frequencies $P(G_{11}|C)$, $P(G_{12}|C)$ and $P(G_{22}|C)$ are calculated p_j^2, $2p_jq_j$ and q_j^2. The trait must be standardised prior to analysis using the population mean and variance, which must either be estimated from the unselected sample or obtained from other sources. The residual trait variance is

$$\sigma_R^2 = 1 - \sum_j P(C_j)(P(G_{11}|C_j)\mu_{11|j}^2 + P(G_{12}|C_j)\mu_{12|j}^2 + P(G_{22}|C_j)\mu_{22|j}^2)$$

where $P(C_j)$ is the prior probability of belonging to class j, calculated by summing posterior probabilities over all N individuals in the sample, $\sum_i P(C_j|G)/N$.

Applying Bayes Theorem to $P(G|T,C)$, the likelihood of observing genotype G_i is the mixture of likelihoods summed over all possible classes weighted by the posterior class probabilities

$$L(G_i|T_i) = \sum_j \frac{P(T_i|G_i,C_j)P(G_i|C_j)}{\sum_F P(T_i|G_F,C_j)P(G_F|C_j)} P(C_j|G_i)$$

where the sum F is over all genotypes. For individual i, the probability of observing the trait score conditional on genotype and class is given by the normal density function

$$P(T_i|G_i,C_j) = \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp -\frac{(T_i - \mu_{G_i|j})^2}{2\sigma_R^2}$$

where G_i is the individual's genotype 11, 12 or 22. The sample log-likelihood is calculated $\sum_i \ln L(G_i|T_i)$. By either fixing or equating parameters, likelihood ratio test statistics can then be constructed between null and alternate models as minus twice the difference in log-likelihood.

It is possible to calculate the class-specific trait means prior to analysis, and to adjust the trait scores accordingly. This is similar to the standard covariate approach advocated in Chapter 6. However, it seems possible that this could cause problems in selected samples (a later section addresses this issue by simulation).

9.2.2 Regression model

As a comparison for the ML method, a standard regression approach is also used in the simulations, with the trait as the dependent variable. As mentioned above, this will not allow correct estimation of the QTL variance, but as regression is typically very robust, it was of interest to investigate further. This procedure is also technically not correct for a further reason: the ML method correctly adopts a weighted mixture of likelihoods to reflect the fact the an individual's class has not necessarily been perfectly measured. That is, the predictor variables are not measured variables but dummy variables that should reflect the uncertainty of class assignment. Although most of the simulations below assume perfect assignment (i.e. all posterior probabilities are 0 or 1) a specific section addresses this issue: the impact of imperfect classification.

For two classes, the additive effects regression model is

$$T_i = \mu + \beta_A A_i + \beta_{C1} C_{1i} + \beta_X A_i C_{1i} + \varepsilon_i$$

where T_i is the trait (mean-centred using the population mean) and A_i and C_{1i} index additive genetic effect (1,0,-1) and class '1' membership (in most cases, either 0 or 1) for individual i. Fixing the β_A coefficient to zero tests for an overall additive effect at the test locus. Fixing β_{C1} to zero tests for a main effect of class. Fixing β_X to zero tests for a QTL $\times$ class interaction. In general, for K classes, $K-1$ 'posterior probabilities' are entered. For example, for three classes the regression equation becomes

$$T_i = \mu + \beta_A A_i + \beta_{C_1} C_{1i} + \beta_{C_2} C_{2i} + \beta_{X_1} A_i C_{1i} + \beta_{X_2} A_i C_{2i} + \varepsilon_i$$

The full list of regression models used in the simulations below are enumerated

here:

$$(1) \quad T_i = \mu + \beta_A A_i + \beta_{C_1} C_{1i} + \beta_{X_1} A_i C_{1i} + \epsilon_i$$

$$(2) \quad T_i = \mu + \beta_A A_i + \beta_{C_1} C_{1i} + \epsilon_i$$

$$(3) \quad T_i = \mu + \beta_{C_1} C_{1i} + \epsilon_i$$

$$(4) \quad T_i = \mu + \beta_A A_i + \epsilon_i$$

$$(5) \quad T_i = \mu + \epsilon_i$$

Several tests can be constructed by comparing the scaled deviance between pairs of regression models. Comparing models (4) and (5) provides a standard test of any QTL effect: this test is susceptible to false positives from population stratification effects. Comparing models (2) and (3) provides a test of association robust to the stratification effects indexed by C. Comparing (1) and (3) also allows for QTL × class interaction. Comparing (1) and (2) specifically tests such an interaction.

The ML method and the regression method therefore handle stratification effects in a slightly different way. The ML method allows allele frequencies to differ between classes whereas the regression method estimates a main effect of class. This reflects the different formulations: as the ML method models genotype conditional on trait, there is no parameter for a strata's main effect on the trait. Conversely, as the regression method models trait conditional on genotype, there is no parameter for allele frequency.

9.2.3 Implementation

The software `L-ASSOC` was developed to implement the ML method. Null and alternate models are specified in terms of the parameters allele frequency p, additive genetic value a and dominance deviation d. Each parameter can be equated across groups (by specifying a lower case letter) or estimated independently within group (by

specifying an upper case letter). Additive genetic values and dominance deviations can be fixed to zero, to provide tests of QTL effect. The data for `L-ASSOC` are the trait score, genotype coded (1, 0 and -1) and posterior probabilities of class membership. The program returns the maximum likelihood parameter estimates and calculates the corresponding components of variance, as well as a likelihood ratio test statistic comparing null and alternate models. The downhill-simplex optimisation method (Nelder and Mead, 1965) was used because of its robustness and simplicity. Very occasionally the algorithm may report a local minimum, but repeating the analysis with different starting values usually resolves this problem.

A test of association (additive and dominance effects) controlling for potential substructure is specified `lassoc --alt Pad --null P` whilst the standard test, ignoring substructure is `lassoc --alt pad --null p`. Allowing different magnitudes of QTL effect between class is specified `lassoc --alt PAD --null P`. To test only additive effect, `lassoc --alt PA --null P`. To test only for differences in allele frequency between classes, `lassoc --alt P --null p`. To test specifically for homogeneity of effects between classes, `lassoc --alt PA --null Pa`. The trait can be adjusted for class-specific mean effects prior to model estimation using the `-m` option.

The regression approach was implemented using the statistical package `R` (a freeware version of `S-Plus`) `lm()` linear model function. For the standard regression test with only a single independent variable the Wald test was used to calculate a χ^2 test statistic, as $(\beta/SE_\beta)^2$. For model comparisons involving more than one parameter, the scaled deviance was used. For normal models, this is $(D_N - D_A)/\hat{\sigma}^2 \sim \chi^2_{df_N - df_A}$ where D_N and D_A are the deviances for the null and alternate model, respectively. The error variance σ^2 is estimated from the residual standard error of the alternate model.

Scheme A	Scheme B	Scheme C
100 lowest	50 lowest	75 lowest
100 highest	150 highest	75 highest
		50 random

Table 9.1: Selection schemes: in all cases, 200 individuals were selected from the entire sample of 1000.

9.3 Simulations: stratified and non-stratified samples

In all of the simulations below, unselected samples of 1000 unrelated individuals measured on a single diallelic test marker and a normally-distributed quantitative trait are randomly generated. Three different extreme sampling schemes are applied to each unselected sample; all selected samples contain 200 individuals. Scheme A selects an equal numbers of individuals from the high and low ends of the trait distribution. Scheme B over-samples individuals at the high end of the distribution, with a smaller group from the low end. Scheme C samples individuals from both ends of the distribution equally, but also includes a number of randomly selected individuals. Table 9.1 summarises these schemes.

All other things being equal, scheme A is likely to be the most efficient selection scheme for the majority of genetic models for the QTL. All schemes are expected to be more efficient than a random sample, however. That is, all schemes select 20% of the sample, but should provide substantially more than 20% of the information for association. Therefore, the informativeness-per-genotype rate is higher, representing a more efficient design.

However, in the presence of stratification, it is possible that selected samples do not retain their desirable characteristics, especially scheme A, which should be the most efficient scheme. Figure 9.1 illustrates trait distributions for two population strata in both unselected and selected samples. A mean difference in trait score between the

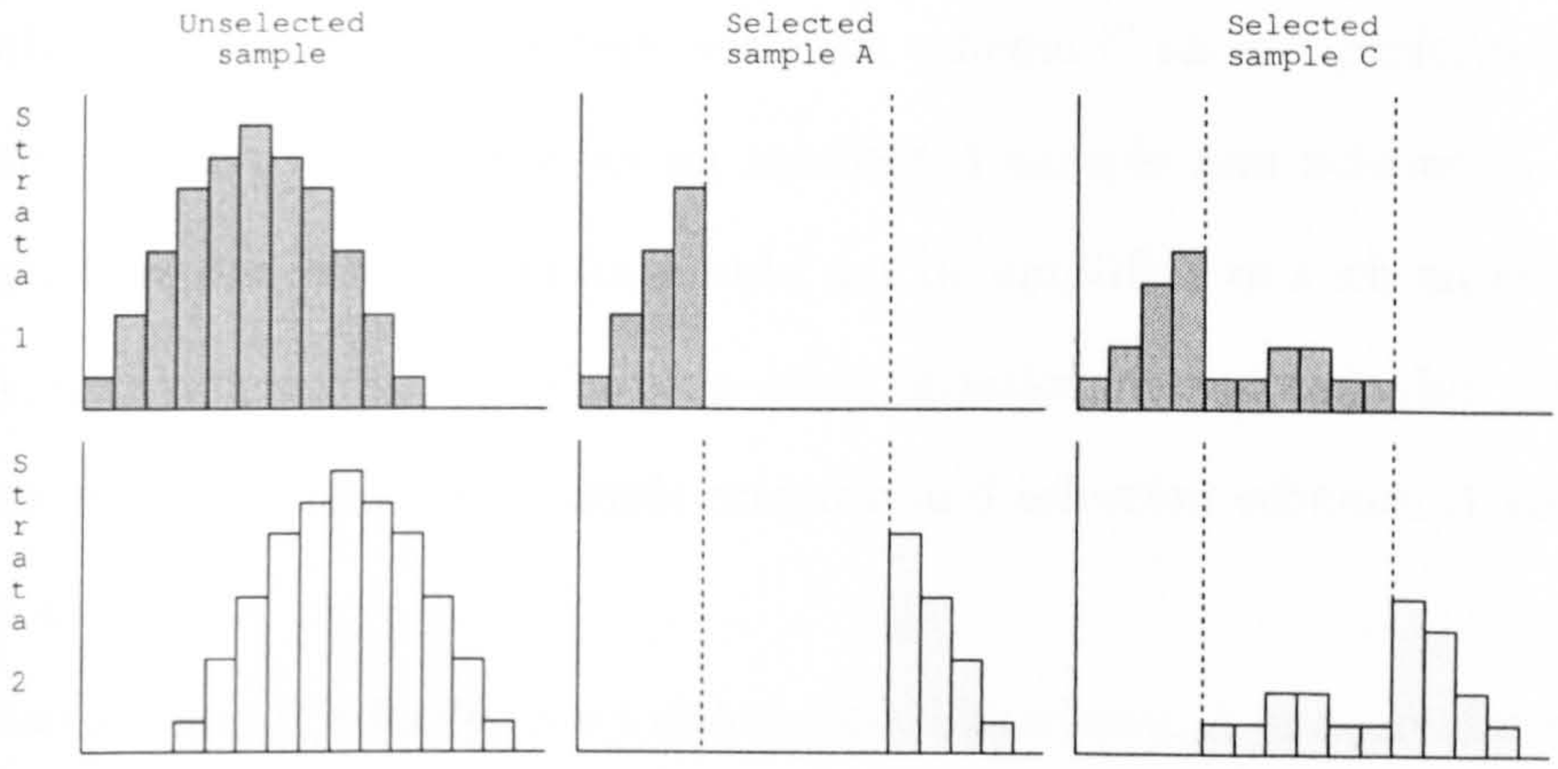


Figure 9.1: Population stratification in unselected and selected samples. See text for details.

two strata is visible in the Figure; in addition, assume a difference in the test locus allele frequency between strata. Two potential problems, which might be expected to hit selection scheme A in particular, can be illustrated with reference to this Figure: (1) enhancing the apparent magnitude of association when stratum membership is unknown and therefore not controlled for, and (2) reducing power to detect the true QTL when stratum membership is known and is controlled for.

That is, in the first instance, where stratum membership is unknown, the top and bottom figures would be merged together into a single, undifferentiated group in analysis. Comparing the unselected sample case with scheme A, the latter essentially amplifies the mean difference between strata. Although this should not increase the probability of detecting a spurious association (as, if there is no true QTL effect at the trait locus, then the stratum allele frequencies will be the same between the unselected and selected samples) it should have the effect of increasing the apparent variance attributable to a true QTL.

More seriously, if, as in the second scenario, stratum membership is known and therefore the association is modelled conditional on stratum membership, then one might expect a reduction in power to detect a true QTL. This is because the within-class variation has been dramatically reduced, and controlling for strata effects is essentially focusing the analysis on within-stratum variation only.

In both scenarios mentioned here, selection scheme C should perform better, as it represents a compromise between an unselected sample and scheme A. That is, any mean differences between strata should not be amplified to such an extent as in scheme A, and a larger degree of within-class variation is retained also. Therefore, a comparison of the unselected sample scenario and selection schemes A and C is of particular interest.

Selection scheme B is likely to act either more like scheme A or more like scheme C depending on the genetic architecture of the test marker, particularly allele frequency. The main focus of the simulations will be only scheme A, however, as well as the comparison between schemes A and C.

Although the Figure represents a very extreme consequence of selection for scheme A (i.e. no strata 1 individuals are in the high group, no strata 2 individuals are in the low group) the effects described above would still be expected to operate in a quantitative fashion with less extreme consequences of selection, e.g. the high group is enriched for individuals from strata 1.

Finally, note that in all cases the test locus is the QTL itself. In other words, the simulations represent candidate gene approaches, rather than mapping strategies which rely on linkage disequilibrium between the test marker and QTL.

9.3.1 Homogeneous samples

The first set of simulations do not generate any between-strata differences – indeed, stratum membership is not even included as a covariate in the analyses. Rather, the aim of these initial homogeneous sample simulations is to assess the performance of the ML and regression methods under the null and the alternate in both unselected and selected samples. The ML method utilises the basic `--alt pa --null p` test; the regression method contrasts regression models (4) and (5).

Two thousand replicates were simulated under each of 15 conditions: the first three

			LRT		Type I error rate	
p	a	d	ML	Reg	ML	Reg
0.5	0	0	0.99	0.99	0.048	0.048
0.1	0	0	0.98	0.98	0.044	0.044
0.9	0	0	0.99	0.99	0.043	0.043

Table 9.2: Homogeneous samples simulated under the null: full samples analysed. Additive genetic effect a and dominance deviation d set to zero, i.e. no QTL effect.

represent the null of no QTL effect, the remaining conditions either an additive-only or additive-and-dominance QTL effect, varying allele frequency and effect size. Results for the full sample under the null are presented in Table 9.2. The allele frequency p is varied but the additive genetic value a and dominance deviation d are always 0. In all cases, results for the ML and the regression method are virtually identical. Over all three conditions, with a nominal type I error rate of 0.05, the average empirical type I error rates were 0.045 for the both ML and regression methods. For the first $p = 0.5$ condition, the empirical type I error rates were very close to 5%, although both methods are possibly slightly conservative with rarer genotypes, i.e. $p = 0.1$ and $p = 0.9$. The expected test statistic is 1 in all cases (i.e. this is a 1 degree of freedom test). Table 9.3 shows the results under the null for selected samples. The average χ^2 test statistics are close to 1 for both ML and regression methods. The average empirical type I error rates are also close to 5% for the ML and regression methods respectively.

For the remaining 12 homogeneous conditions simulated under the alternate hypothesis (i.e. with a QTL effect), Table 9.4 gives the expected additive (σ^2_A) and dominance (σ^2_D) variance components. As can be seen, the final two conditions (representing a common allele with a dominant effect) are unlikely to be detectable (i.e. equivalently, this represent a rare recessive effect). Otherwise, the QTL typically accounts for between 1% and 10% of the trait variance.

Table 9.5 shows results under the alternate models for the basic test of association

			LRT		Type I error rate	
p	a	d	ML	Reg	ML	Reg
Scheme A						
0.5	0	0	1.00	1.01	0.051	0.052
0.1	0	0	1.02	1.01	0.052	0.051
0.9	0	0	0.99	0.98	0.046	0.046
Scheme B						
0.5	0	0	0.99	0.99	0.051	0.052
0.1	0	0	1.02	1.02	0.057	0.055
0.9	0	0	1.02	1.01	0.049	0.050
Scheme C						
0.5	0	0	1.05	1.06	0.053	0.055
0.1	0	0	1.04	1.03	0.058	0.058
0.9	0	0	1.04	1.04	0.056	0.054

Table 9.3: Homogeneous samples simulated under the null: selected samples analysed.

not controlling for stratification. In all but the last two conditions, the χ^2 test statistics are highly significant (all are 1 degree of freedom tests). In every case, the regression approach has slightly larger test statistic values than the ML method, although this difference is barely noticeable for test statistics of small to moderate size.

The values tabulated for the three selected sampling schemes represent the average test statistic for the selected sample as a proportion of the average test statistic for the entire sample. The sampling schemes typically retain around 65%, 55% and 60% of the information, for schemes A, B and C respectively, for both ML and regression methods. That is, despite only retaining 20% of the sample, considerably more of the information for association is retained. Selection scheme A is confirmed as the most efficient design. The ML and regression approaches appear to have roughly similar profiles with respect to their behaviour in selected samples. One minor difference appears to be for the asymmetrical scheme B when $p = 0.1$ – the ML approach retains relatively more of the full sample information in these cases (and in most cases, the absolute value of test statistics is higher too).

	p	a	d	σ^2_A	σ^2_D
Additive	0.5	0.25	0	0.03	0.00
	0.5	0.5	0	0.11	0.00
	0.1	0.25	0	0.01	0.00
	0.1	0.5	0	0.04	0.00
	0.9	0.25	0	0.01	0.00
	0.9	0.5	0	0.04	0.00
With dominance	0.5	0.25	0.25	0.03	0.01
	0.5	0.5	0.5	0.11	0.05
	0.1	0.25	0.25	0.04	0.00
	0.1	0.5	0.5	0.13	0.01
	0.9	0.25	0.25	0.00	0.00
	0.9	0.5	0.5	0.00	0.01

Table 9.4: Parameters and expected additive and dominance variance components used to simulate homogeneous samples under the alternate hypothesis of a QTL effect.

			Full Sample		Scheme A		Scheme B		Scheme C	
p	a	d	ML	Reg	ML^1	Reg^1	ML^1	Reg^1	ML^1	Reg^1
0.5	0.25	0	31.48	31.98	0.65	0.68	0.54	0.55	0.60	0.62
0.5	0.5	0	118.69	126.03	0.62	0.72	0.50	0.57	0.56	0.63
0.1	0.25	0	12.40	12.47	0.68	0.67	0.54	0.47	0.59	0.59
0.1	0.5	0	45.09	46.17	0.62	0.62	0.45	0.37	0.57	0.56
0.9	0.25	0	11.85	11.92	0.67	0.66	0.60	0.69	0.63	0.62
0.9	0.5	0	45.26	46.37	0.62	0.61	0.59	0.75	0.57	0.57
0.5	0.25	0.25	31.35	31.85	0.64	0.68	0.59	0.65	0.58	0.61
0.5	0.5	0.5	111.20	117.64	0.63	0.73	0.60	0.76	0.56	0.64
0.1	0.25	0.25	36.09	36.76	0.64	0.65	0.48	0.40	0.59	0.60
0.1	0.5	0.5	135.30	145.87	0.57	0.64	0.40	0.32	0.53	0.59
0.9	0.25	0.25	1.45	1.45	0.90	0.87	0.95	0.99	0.94	0.92
0.9	0.5	0.5	2.97	2.98	0.84	0.77	0.98	1.03	0.80	0.75

Table 9.5: Homogeneous samples simulated with a QTL effect: a comparison of ML and regression approaches, in full and selected samples. For the full sample the actual χ^2 test statistics are shown. For the selected samples, ML^1 and Reg^1 represent the proportion of the full sample test statistic obtained.

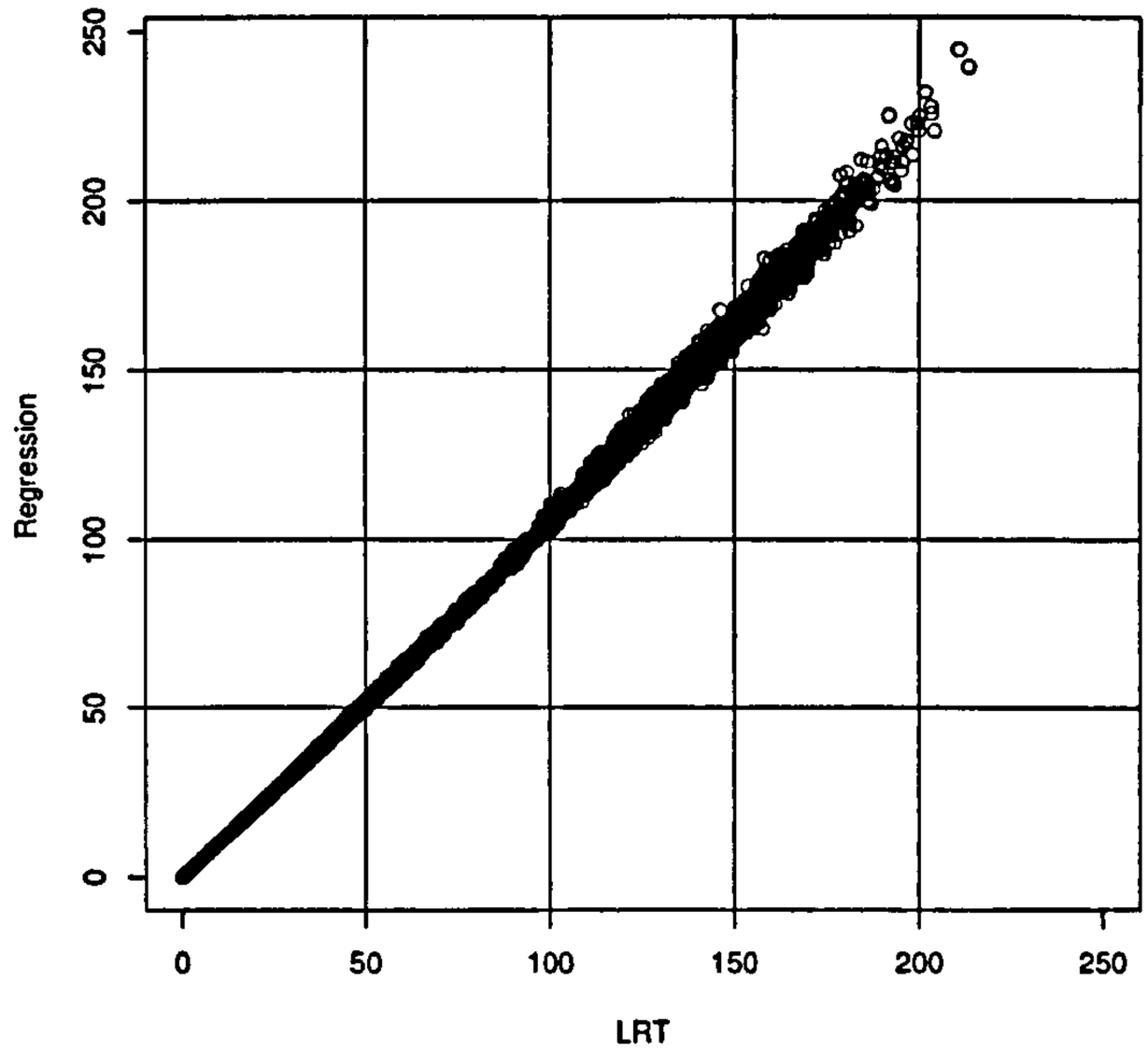


Figure 9.2: A comparison of the ML (LRT) and regression approaches: test statistics from the standard additive models, in full samples. At higher values, the regression approach gains a greater advantage over the ML approach.

From the additive-only simulations, Figure 9.2 plots the test statistics of the ML method against those of the regression method, over all 15 conditions. In general, both methods are roughly equivalent. For very large values (> 100) the regression approach becomes more powerful – there is a slight upwards curve in the band of points. For realistic effect sizes, both methods are likely to be comparable, however. In terms of statistical power, rather than test statistic value, power will reach a plateau near 100% for such large test statistic values for both ML and regression methods.

To summarise the results of the homogeneous (no-stratification) conditions: both ML and regression methods appear to perform well under the null and under the alternative in both selected and unselected samples. If anything, the regression method appears to be more powerful.

9.3.2 Heterogeneous samples

The main emphasis of this Chapter is to look at the properties of selected samples when stratification is present. In these circumstances, stratification effects may or may not be explicitly modelled in analysis. The simulations below look at the impact of modelled and unmodelled stratification in unselected and selected samples on ML and regression approaches to QTL association.

In all cases, only two strata are simulated. Unless otherwise specified, the strata represent an even 50:50 split of the entire sample. Furthermore, unless otherwise specified it is assumed that stratum membership is perfectly measured. That is, all 'posterior probabilities' are either 0 or 1. Table 9.6 gives an overview of the conditions simulated in this section. The numbers '*1*' to '*5*' indicate the presence or absence of between strata differences in trait mean and/or allele frequency.

Only under conditions '*4*' and '*5*' would one expect spurious association results when stratum membership is not controlled for. Furthermore, only condition '*4*' would be expected to reduce power to detect a true QTL effect if stratum membership is not controlled for – this is referred to as 'masking' stratification, where the true QTL increaser allele is less frequent in the class with the higher mean. Conversely, condition '*5*' would amplify the test statistic associated with a true QTL. For conditions with no allele frequency differences ('*1*' and '*3*') the QTL allele frequency is 0.5 in both strata. For the other conditions, the QTL frequency is 0.4 in one stratum, 0.6 in the other. This corresponds to a 'Delta' value of 0.2, as used in many of the simulations in Chapter 6.

The letters '*a*' to '*e*' represent the model used to simulate the actual QTL. Additive effects were set with an additive genetic value $a = 0.5$; if there is also dominance then $a = d = 0.5$. Models '*b*' and '*d*' represent QTL × class interactions, for which the presence of the QTL effect depends upon stratum membership.

In addition, a residual component of variation (with a variance of 1) was added

Set	Description
1	No class differences
2	QTL allele frequency differences only
3	Mean differences only
4	Masking stratification (increaser allele less frequent in class with higher mean)
5	Normal stratification (increaser allele more frequent in class with higher mean)
a	No QTL effects
b	Additive effects in one class only
c	Additive effects in both classes
d	Additive and dominance effects in one class only
e	Additive and dominance effects in both classes

Table 9.6: Heterogeneous simulations: summary of conditions.

to the trait. The exact proportion of variance attributable to the QTL will, however, depend on other factors. For example, a mean difference between the two strata, if unmodelled, effectively increases the residual variance and so decreases the proportion of variance explained by the QTL. That is, one is not necessarily comparing like with like when looking across the different conditions. These incidental effects have no major impact on the results however – we are more interested in comparing across table columns, i.e. methods and samples, than table rows, i.e. conditions.

Combining the five 'stratification models' with the five 'QTL models' gives 25 conditions. Initially, we shall only consider the additive models, '*a*' to '*c*', giving 15 conditions. Table 9.7 presents the results for these conditions: for the full sample and samples selected under scheme A, the average test statistics from 500 replicates are presented for ML and regression approaches.

Table 9.7 is the most important table in this Chapter, illustrating several properties of the methods as applied to selected samples under stratification. Each row indicates the results for one of the 15 additive-only conditions, as indicated in the first column. The next four columns give the results for the full sample: comparing ML and regression test statistics when stratum membership is not modelled (columns 2 and 3) and then comparing the two methods when stratum membership is modelled

	Full sample				Scheme A			
	No covariate or interaction		Covariate & interaction		No covariate or interaction		Covariate & interaction	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg
1a	1.05	1.10	1.93	2.15	1.09	1.20	2.04	2.31
2a	1.03	1.06	2.05	2.13	1.05	1.02	2.01	1.94
3a	1.00	1.06	2.08	1.78	1.01	1.17	2.23	2.00
4a	40.31	42.79	2.14	2.10	15.96	17.02	2.19	2.12
5a	40.85	41.69	2.00	1.87	16.15	16.65	2.24	1.84
1b	31.75	31.57	61.70	65.02	21.31	20.47	40.08	40.44
2b	28.32	28.77	59.15	62.64	18.99	20.13	38.13	39.98
3b	16.08	16.47	60.80	64.26	13.74	14.98	7.87	9.75
4b	7.19	7.11	57.75	62.23	1.06	1.09	7.45	8.38
5b	100.67	106.55	57.88	60.94	53.80	62.22	8.48	9.89
1c	119.59	126.43	120.35	127.39	73.51	91.90	74.27	92.16
2c	106.04	113.21	116.04	124.98	68.05	81.57	72.06	88.64
3c	62.09	62.24	118.42	126.98	52.61	57.93	12.54	13.80
4c	2.19	2.10	113.13	121.47	10.76	11.86	12.43	12.46
5c	191.35	211.25	111.43	121.36	112.34	152.37	12.28	15.77

Table 9.7: Heterogeneous simulations: main results for full sample and scheme A, applying ML and regression approaches to additive-only models.

(columns 4 and 5). The next four columns provide similar figures, but for samples selected under scheme A rather than the full unselected sample.

The "No covariate, no interaction" ML model is specified `--alt pa --null p` as before; alternatively comparing regression models (4) and (5) provides a conceptually equivalent test. This test has 1 degree of freedom, so the expected test statistic under the null is also 1. The "Covariate and interaction" ML model is specified `--alt PA --null P`; comparing regression models (1) and (3) provides a conceptually equivalent test. For two strata, this test has 2 degrees of freedom, so the expected test statistic under the null is 2.

The 15 conditions are ordered in 5 groups of 3: the three QTL conditions are '*a*' no QTL effect, '*b*' an additive-only QTL effect in one class or '*c*' an additive-only QTL effect in both classes. In general, the regression method appears to perform better than the ML method – both under the null and the alternate hypothesis, as seen

in the previous simulations. Unless otherwise mentioned, however, the discussion of results below applies equally to ML and regression approaches.

Full sample results from 9.7

The first five rows of Table 9.7 represent conditions where no QTL effect was simulated. For the full sample, the results clearly show that spurious association occurs under conditions 4*a* and 5*a* when stratum membership is not modelled. If there are only allele frequency differences (2*a*) or trait mean differences (3*a*) between strata, then spurious association does not result (i.e. the test statistics are near their expected value of 1 even when stratum membership is not modelled). However, when stratum membership is modelled, the test statistics return to near their expected value under the null for 4*a* and 5*a*.

When a QTL effect is present, the 'masking' and 'amplifying' effects of unmodelled stratification can be seen in the full sample results for 4*b* and 4*c* (masking) and 5*b* and 5*c* (amplifying). When stratification is modelled in the full sample, however, the masking and amplifying effects disappear, as expected.

As mentioned above, whether or not a mean trait difference is simulated has an impact on the effective proportion of variance the QTL explains. This is reflected in the test statistics: for example, for the ML method, the average test statistic for 3*c* is 62.09 compared to 119.59 for 1*c* – this reduction is solely due to the smaller effective QTL effect rather than any 'masking' effect, however. As expected, when stratum membership is modelled, this attenuation disappears (conditional on stratum, the effective QTL variance is now the same in condition '*3*' as '*1*') and the test statistics are 118.42 and 120.35 respectively.

For the full sample when strata membership is modelled, the test statistics for the two stratification conditions '*4*' and '*5*' are similar to those for the conditions '*1*' to '*3*'. That is, controlling for the stratification when it is present doesn't drastically

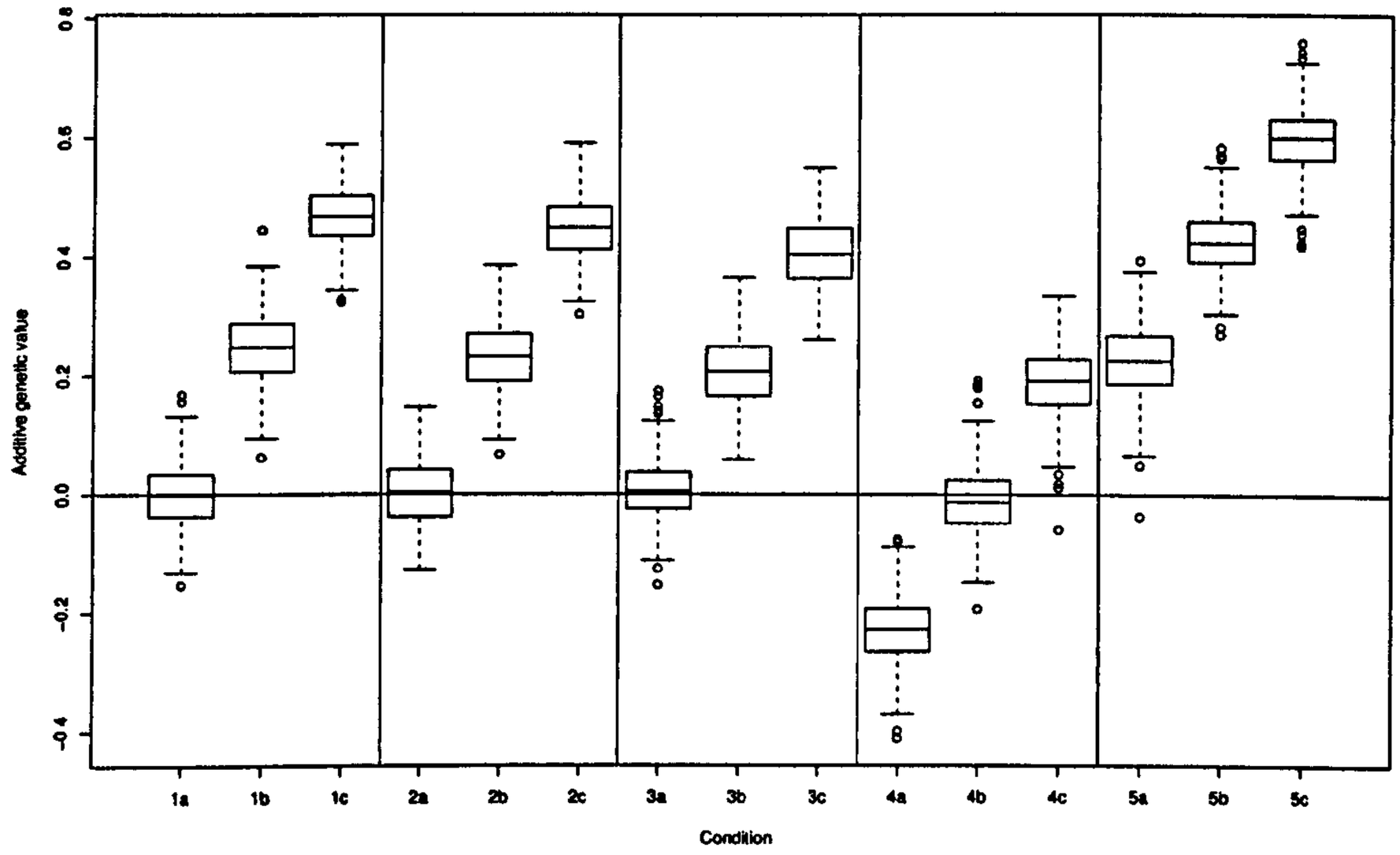


Figure 9.3: Boxplots of the estimated additive genetic values.

reduce the power to detect a true QTL effect. (Actually, there is a slight reduction, e.g. for the '*c*' conditions, the ML statistics for '*4*' and '*5*' are around 112 whereas for '*1*' to '*3*' they are around 118.)

Figure 9.3 plots the estimated additive genetic values for the 15 conditions from the ML method applied to the full sample not controlling for stratification – the impact of masking and amplifying effects are clearly visible.

Scheme A results from Table 9.7

Considering now the results for samples selected under scheme A: the simulations under the null of no QTL effect, condition '*a*', in the selected sample results are similar to the full sample results. As expected, there is less evidence for spurious association in conditions '*4*' and '*5*' than for unselected samples, due to the smaller sample size (i.e. around 16 as opposed to 40).

There are similar masking and amplifying effects in the presence of unmodelled stratification on the power to detect a true QTL. However, a major difference between

the unselected sample results and scheme A results arises when looking at cases where a stratification effect is modelled. For conditions '*4*' and '*5*', but also for '*3*' as well, there is a marked attenuation in the test statistics when there is a true QTL effect ('*b*' and '*c*' conditions) when stratification effects are modelled. For example, consider the *1c* condition where there is no real stratification. Contrasting the ML results for the tests that model potential strata effects in the unselected and selected samples we can see that sample selection at a 20% level retains 74.27/120.35 = 62% of the information for association. However, for a condition where there is stratification, such as 5*c*, sample selection at the 20% level retains only 12.28/111.43 = 11% of the information for association. Selected samples typically represent more efficient designs because the proportion of the sample selected will be smaller than the proportion of the information retained – this is not the case when one attempts to control for strata effects in a selected sample, however. Note that this effect occurs even if there is only a mean difference between strata (i.e. condition *3*). This occurs for the reasons outlined above – the reduction in within-stratum trait variation that can arise from selecting extremes from a heterogeneous sample. Similar results are obtained from the regression statistics also.

As suggested earlier, it is possible that sample selection might artificially inflate the estimated variance attributable to a QTL under stratification. Table 9.8 provides some weak support for this phenomena, showing the average estimated variance component under the basic pa model. The QTL variance does appear to be greater in the selected samples for conditions '3*b*' and '3*c*' (where there is a mean difference only between strata) although the results are not clear for the stratification conditions '*4*' and '*5*'. None of the differences are particularly striking, however, although this may change for different selection, stratification and/or QTL conditions.

In summary, the results of Table 9.7 indicate the following: (1) for both ML and regression methods, unmodelled stratification can induce spurious association

	Estimated proportion of trait variance	
	Full	Scheme A
1a	0.001	0.001
2a	0.001	0.001
3a	0.001	0.001
4a	0.038	0.028
5a	0.038	0.027
1b	0.031	0.032
2b	0.026	0.029
3b	0.016	0.023
4b	0.007	0.002
5b	0.094	0.090
1c	0.110	0.111
2c	0.096	0.102
3c	0.059	0.085
4c	0.002	0.020
5c	0.170	0.180

Table 9.8: Heterogeneous simulations: estimated proportion of variance explained by the QTL under the basic pa model.

as well as masking or amplifying the effects of a true QTL, in both unselected and selected samples (2) these phenomena can be can controlled for, within both the ML and regression approaches, by explicitly modelling the strata effects (3) in general the regression approach slightly out-performs the ML approach (4) modelling strata effects in selected samples can lead to a serious attenuation in power to detect true QTL.

9.3.3 Modelling dominance

The following simulations confirm that dominance effects can be modelled correctly using the ML method (the regression method could also model dominance, by coding a further independent variable as mentioned at the beginning of this Chapter). For homogeneous samples, the correct type I error rates are obtained under the null with the test `--alt pad --null p`. The average test statistic is 2.033 (expected value of 2) and the average empirical type I error rate is 0.055 (expected 0.05).

	Full sample		Scheme A	
	LRT	*p-value*	LRT	*p-value*
1a	2.07	0.486	2.12	0.481
2a	2.01	0.505	2.11	0.492
3a	2.22	0.485	1.92	0.505
4a	2.11	0.484	1.91	0.510
5a	2.07	0.505	1.94	0.512
1b	1.91	0.502	1.95	0.501
2b	1.91	0.512	1.91	0.509
3b	2.70	0.418	2.23	0.473
4b	2.28	0.464	2.07	0.489
5b	2.56	0.425	2.06	0.485
1c	2.03	0.498	2.04	0.506
2c	2.10	0.485	2.04	0.498
3c	2.88	0.394	2.26	0.486
4c	2.67	0.423	2.33	0.457
5c	2.82	0.418	2.24	0.458
1d	32.01	0.000	17.30	0.011
2d	29.41	0.001	15.25	0.021
3d	15.78	0.019	6.73	0.150
4d	16.70	0.011	5.91	0.171
5d	12.70	0.040	7.01	0.148
1e	61.70	0.000	29.79	0.000
2e	56.24	0.000	27.64	0.001
3e	39.74	0.000	17.97	0.008
4e	38.85	0.000	19.07	0.006
5e	36.14	0.000	16.09	0.012

Table 9.9: Modelling dominance effects: heterogeneous samples. The likelihood ratio test compares model PAD against PA which is, for 2 classes, a 2 degree of freedom test.

Table 9.9 gives results for specific tests of dominance effects over and above pure additive effects, in heterogeneous samples when stratification effects may be present. Only QTL models '*d*' and '*e*' have dominance effects. The test is specified as `--alt PAD --null PA` which is a 2 degree of freedom test (i.e. it allows for different effects between groups). For the full sample, the average *p*-value is below the critical value of 0.05 in all '*d*' and '*e*' cases, although power to detect specific dominance effects is lower when stratification effects are present (e.g. 5*d*). In selected samples, all '*e*' models show significant specific dominance effects (i.e. where both strata have dominance

	Scheme B				Scheme C			
	No covariate or interaction		Covariate & interaction		No covariate or interaction		Covariate & interaction	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg
1a	1.06	1.20	1.96	2.49	1.15	1.10	2.24	1.95
2a	1.07	1.01	2.03	1.92	1.12	1.08	2.17	2.13
3a	1.01	1.19	2.26	2.68	0.85	0.89	2.11	2.01
4a	11.92	12.57	2.32	2.43	13.97	15.03	2.31	2.09
5a	12.04	12.33	2.37	2.62	14.05	14.66	2.03	1.94
1b	18.14	16.69	33.35	34.76	19.27	18.75	36.36	36.67
2b	16.36	17.12	30.74	31.51	17.28	16.83	35.47	34.99
3b	9.05	9.98	7.30	6.21	12.52	12.64	13.99	15.46
4b	1.24	1.39	7.76	6.39	1.09	1.12	13.61	15.38
5b	38.12	44.25	7.28	5.58	48.09	54.31	13.85	16.23
1c	59.44	72.57	60.43	73.29	66.44	78.56	67.25	78.97
2c	54.68	65.04	58.34	71.15	60.36	71.17	64.54	76.86
3c	43.37	47.32	15.46	28.22	46.07	52.20	25.41	28.27
4c	10.70	11.10	14.96	23.88	10.64	11.41	24.35	26.36
5c	90.22	117.23	15.30	32.31	99.52	128.42	24.44	28.62

Table 9.10: Heterogeneous simulations: main results for alternative selected sampling schemes B and C.

effects); for '*d*' models, mean differences between strata appear to interfere ($3d - 5d$).

9.3.4 Alternate selected sampling schemes

As mentioned above, it might be expected that selection scheme A is particularly vulnerable to the reduction in power effect observed in the presence of stratification. This section considers the scheme B and C results for the heterogeneous additive-only models (Table 9.10).

As expected, scheme B shows a marginally smaller impact of unmodelled stratification ($4a$ and $5a$) compared to schemes A and C, reflecting the fact that it is, under most circumstances, a less efficient design (i.e. a design with less power to detect true association will also be less affected by spurious association). Scheme C is affected to roughly the same extent as scheme A.

For scheme B, the test statistics under the null (condition '*a*') when modelling

stratification appear to be slightly too high, particularly for the regression method. The expected value is 2, whilst the ML average over the 5 stratification conditions is 2.188, the regression average is 2.428. For scheme C, the results are closer to the expected value (2.172 and 2.024 for ML and regression methods respectively).

Focusing on the results for the '*c*' conditions (an additive QTL operating in both strata), for which the results are clearest and most pronounced, some interesting differences emerge. Firstly, as predicted, scheme C does not show the same reduction in power as scheme A when modelling stratification effects. For example, taking 5*c*, scheme C retains 24.44/111.43 = 22% of the information for association using the ML method (scheme A retained only 11%). Performance is equivalent to a randomly selected subsample of 20% therefore, but no worse. This ratio is similar for the regression method.

Scheme B shows a differential reduction in power between ML and regression approaches – the regression approach is less affected (retaining 32.31/121.36 = 27%) than the ML approach (15.30/111.43 =14%). To what extent this result generalises is not immediately clear. For example, for the '*b*' conditions, this pattern is not seen (in fact, the ML method is marginally better).

In general, however, these results support the idea that a sampling scheme such as C would perform better in the presence of modelled stratification whilst still being almost as efficient as scheme A under homogeneous conditions.

It is also worth noting that in all cases the selected samples constitute quite a large proportion of the entire sample (20%). Presumably, the reduction in power would grow worse with more extreme sampling schemes. Of course, ideally a correction for stratum membership would be made prior to sample selection. This may not always be possible, however, especially if stratum membership is estimated via a genetic background approach.

	Full sample		Scheme A		Scheme B		Scheme C	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg
1a	1.05	1.10	1.10	1.19	1.05	1.19	1.14	1.11
2a	1.01	1.10	0.98	1.02	1.01	0.99	1.07	1.06
3a	1.01	0.95	1.07	0.94	1.05	0.93	1.09	0.96
4a	1.11	1.14	1.18	0.98	1.10	0.98	1.13	1.14
5a	0.90	0.94	1.02	0.87	1.08	0.78	0.80	0.97
1b	31.72	31.56	21.27	20.47	18.55	17.01	19.25	18.65
2b	30.38	30.99	19.78	21.43	16.25	17.17	18.27	18.10
3b	31.16	31.49	4.56	4.66	3.55	3.65	6.99	7.64
4b	29.82	28.89	4.47	4.04	4.47	3.71	7.39	7.53
5b	29.57	29.40	4.90	4.23	3.63	3.26	6.92	7.35
1c	119.40	126.32	73.29	91.63	59.40	72.26	66.37	78.25
2c	114.93	124.04	70.95	88.04	57.26	70.04	63.59	76.37
3c	117.41	125.69	11.58	12.10	14.41	14.78	24.48	27.13
4c	112.16	120.55	11.56	11.06	13.96	13.63	23.26	25.24
5c	110.75	120.44	11.19	13.57	14.21	16.25	23.33	27.25

Table 9.11: Controlling for main effects of stratification only. The ML tests compares models Pa and P. The regression test compares models (2) and (3).

9.3.5 Modelling class-specific means only

Previously, all of the models that have taken class structure into account have allowed for interactions between genetic effects and class. That is, they have allowed for an effect being present in only one of two classes (i.e. the '*b*' condition). Of course, it is possible to constrain the models such that any additive genetic (or dominance) effect is constant across all strata. The ML method still allows allele frequency to vary between strata; the regression approach still includes a term to estimate any stratum-specific mean effects. The ML model is specified as `--alt Pa --null P`. The regression test is formed by comparing regression models (2) and (3). Table 9.11 shows the results for the additive-only models.

The basic result is that, as expected, the statistics for '*a*' and '*c*' conditions remain unchanged (i.e. there was no QTL × class interaction in those cases). The '*b*' condition results are greatly attenuated relative to the previous tests that allowed for QTL × class interaction.

	ML				Reg			
	Full	A	B	C	Full	A	B	C
5a	1.10	1.23	1.29	1.23	0.93	0.97	1.83	0.97
5b	28.16	3.59	3.65	6.93	30.66	5.52	2.31	8.55
5c	0.90	1.14	1.02	1.11	0.93	2.11	14.58	1.33

Table 9.12: QTL × class interaction: ML versus regression approaches.

9.3.6 Specific tests of QTL × class interaction

As well as testing only for a main effect of strata, it is possible to perform a specific test of QTL × strata interaction, in the presence of any main strata effects. That is, the ML model is specified `--alt PA --null Pa`. The regression test compares equations (1) and (2). This test has an expected χ^2 of 1 under the null.

Table 9.12 shows some of the results for conditions 5*a*, 5*b* and 5*c*. The results for stratification conditions '*1*' – '*4*' were largely as expected, with the '*b*' conditions showing significant test results, the '*a*' and '*c*' conditions not doing so. In the presence of stratification, as in '*5*', the ML approach behaves as expected, both under unselected and selected samples. The regression approach shows some problems in selected samples. For example, for 5*c* the scheme B statistic is grossly inflated at 14.58.

9.3.7 Specific tests of allele frequency differences

A test for allele frequency differences between classes can be constructed within the ML framework by specifying the alternate model as `--alt P` and the null model as `--null p`. Table 9.13 shows the results, ordered by stratification condition rather than QTL condition: '*1*' no strata effects, '*2*' allele frequency difference only, '*3*' mean difference only, '*4*' masking stratification, '*5*' normal stratification. When there are no strata differences at all, the test statistics are near their expected value in both unselected and selected samples. When there are allele frequency differences, the test statistics are much larger (around 80 in unselected samples, 16 in selected samples).

	Full sample	Scheme A
1a	1.09	1.10
1b	0.96	0.98
1c	0.98	1.08
2a	82.04	17.31
2b	82.71	15.33
2c	81.71	14.12
3a	1.08	0.99
3b	1.04	11.15
3c	1.26	43.18
4a	81.04	17.19
4b	82.43	1.62
4c	82.97	6.22
5a	83.75	17.36
5b	80.48	50.07
5c	82.51	100.80

Table 9.13: Likelihood ratio test statistics for specific tests of allele frequency differences between classes.

However, when there is a mean difference (but no allele frequency difference) between strata, the presence of a QTL effect in selected samples is interpreted as an allele frequency difference. In this case, it might be desirable to specify the test as `--alt Pa --null pa`.

9.3.8 Unequal subpopulations sizes

All previous simulations have had two classes occurring at equal frequencies in the unselected sample – 500 individuals in each class. This section looks at the effect of unequal strata frequencies – in this case, a 1:9 split.

Repeating the heterogeneous simulations with unequal strata sizes, Table 9.14 shows a subset of results (only for no QTL effects or additive QTL effects in both strata, i.e. '*a*' or '*c*'). Unsurprisingly, the impact of stratification is now not as severe under these conditions compared to the 50:50.

	Full sample				Scheme A			
	No covariate or interaction		Covariate & interaction		No covariate or interaction		Covariate & interaction	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg
1a	0.94	0.93	2.01	1.98	0.87	0.86	1.91	1.93
2a	0.92	0.92	2.03	2.01	0.93	0.94	2.03	2.07
3a	1.00	1.00	2.02	2.00	1.00	0.99	1.53	2.18
4a	9.10	9.16	1.94	1.92	8.09	8.00	1.49	1.97
5a	8.86	8.92	1.95	1.94	8.01	7.96	1.72	2.29
1c	119.17	126.58	120.19	127.41	90.73	73.61	90.74	74.57
2c	110.43	116.78	115.28	121.99	82.79	69.16	85.68	71.61
3c	89.26	93.16	120.14	127.69	59.70	51.77	82.81	54.83
4c	39.95	40.51	113.44	120.15	20.72	20.05	70.40	50.05
5c	144.66	156.56	115.94	123.13	115.29	90.86	85.92	54.33

Table 9.14: Heterogeneous simulations with unequal simulated class sizes: ML versus regression approaches.

The spurious association due to unmodelled stratification is less in this case (e.g. 8.86 compared for 40.85 for 5*a* in unselected samples). Likewise, the reduction in power due to modelling stratification effects in selected samples is less when strata sizes are unequal. In general, choosing a 50:50 mixing proportion represents a 'worst case scenario' of stratification.

Whether or not there was a 100:900 or a 900:100 mixture did not effect the results under these conditions. Of course, this might not always be the case, e.g. if the QTL has unequal allele frequencies and there is a mean difference between strata, possibly.

9.3.9 Impact of sample outliers

As the ML approach models the genotype conditional on trait values, it should be more robust to outlying trait values than the regression method which has the trait as the dependent variable. A cubic transformation was applied to the trait distribution before standardising it, which has the effect of producing extreme population outliers. As Table 9.15 shows, under these circumstances the ML approach does indeed often perform marginally better than the regression approach. Otherwise, all the familiar

	Full sample				Scheme A			
	No covariate or interaction		Covariate & interaction		No covariate or interaction		Covariate & interaction	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg
1a	0.95	0.94	2.02	1.99	0.94	0.94	2.02	1.99
2a	0.95	0.95	2.01	1.99	0.99	0.96	2.02	2.01
3a	0.96	0.96	2.10	2.08	0.94	0.94	2.01	1.95
4a	17.69	17.37	1.97	1.93	11.58	11.48	2.20	2.11
5a	17.92	17.59	1.91	1.89	11.69	11.66	2.07	2.00
1b	21.25	20.24	41.10	40.40	17.39	16.88	33.03	31.33
2b	18.69	17.90	39.17	39.13	16.01	15.12	31.49	30.05
3b	15.97	15.61	39.26	39.73	12.92	13.09	6.30	6.20
4b	1.04	1.04	38.44	41.78	1.02	1.03	6.12	6.21
5b	62.21	59.13	37.51	34.74	44.51	45.06	6.05	5.70
1c	77.03	68.21	78.71	69.42	59.86	59.03	61.47	59.77
2c	67.21	60.84	75.99	67.86	55.01	53.37	59.20	58.37
3c	56.89	54.01	74.70	69.69	47.08	49.79	10.03	9.05
4c	10.83	10.67	74.09	75.14	13.41	14.10	9.97	9.53
5c	134.73	124.05	73.55	62.90	99.68	106.26	10.09	8.61

Table 9.15: Impact of population outliers on ML and regression approaches.

signs of stratification can be observed.

9.3.10 Correcting for subpopulation mean effects

As mentioned, although it is not possible to estimate strata means in the ML approach, it is possible to calculate them from the sample and adjust the scores accordingly. This option is specified by the `-m` option, which appropriately weights the scores by the posterior class probabilities when calculating the class means. Table 9.16 shows the results from the basic `--alt pa --null p` model after applying the mean-correction for both unselected and selected samples A, B and C.

This approach ensures valid test statistics under the null even in the presence of stratification – that is, 4*a* and 5*a* do not show inflated values, even in selected samples. This approach also largely removes the masking/amplifying effects of unmodelled stratification in the full sample. However, the reduction in power in selected samples

	Full	A	B	C
1a	1.05	0.95	0.94	1.13
2a	0.92	0.95	0.89	0.91
3a	0.93	1.13	1.13	0.97
4a	0.93	0.91	0.84	0.96
5a	0.89	0.85	0.87	0.91
1b	31.64	21.32	18.03	21.26
2b	26.69	18.35	15.00	19.44
3b	30.87	3.97	3.54	6.96
4b	26.51	4.19	4.11	6.83
5b	26.18	2.80	2.43	5.61
1c	119.09	73.42	59.40	73.87
2c	104.42	67.28	54.45	66.40
3c	115.58	8.33	11.26	23.70
4c	99.47	11.31	13.97	26.33
5c	99.69	4.91	8.50	17.79

Table 9.16: Results for standard ML test after correcting for class-specific mean effects; for the full sample and the three selected sampling schemes.

appears even worse using this approach. For example, comparing scheme A with the full sample for 5*c*, only 5% (4.91/99.69) of the information is retained by 20% of the sample (compared to 11% when the full `pa/p` model was applied).

It is not immediately clear what causes this pattern of results: the approach of mean-correcting is analogous to the way in which the regression approach models the main effects (i.e. which is equivalent to performing the regression analysis on the trait residual after the effect of strata has been partialled out).

9.3.11 Imperfect classification

For the initial heterogenous simulations we assume that stratum membership is perfectly measured. That is, individuals' posterior probabilities are only ever 0 or 1. In practice, of course, this will not always be the case, especially if genetic background methods are used to estimate stratum membership. Although it might be possible to achieve near-perfect classification, if, for example, enough markers are used in the genetic background procedure and the population substructure is clear, one might

expect less than perfect classification in many instances. This section examines the impact of different types of imperfect stratum membership assignment.

For data simulated under the heterogeneous conditions, the posterior probabilities for each individual are changed in one of three ways. The first condition introduces uncertainty by changing all 1 values to 0.8 and all 0 values to 0.2. That is, [1,0] becomes [0.8,0.2] and [0,1] becomes [0.2,0.8]. The second condition introduces misclassification, such that a random 20% of individuals have their posterior probabilities switched, where [1,0] becomes [0,1] and [0,1] becomes [1,0]. In this way, the overall mixing proportions of the two classes will remain constant, i.e. 50:50, for both the first and second conditions. The third condition mimics a more realistic pattern of classification. The probabilities are sampled from $N(0.2, 0.1)$ and $N(0.8, 0.1)$ distributions, bounded at 0 and 1. Figure 9.4 illustrates a typical distribution of probabilities. The two modes represent the two strata – whilst there is clear separation of these two peaks, not all individuals can be unambiguously classified. One might expect such a distribution of posterior probabilities if the genetic background analysis is somewhat under-powered.

For models '*1*' – '*3*' (i.e. no stratification) there is little or no impact from imperfect classification. This is obvious, as the class structure will have no relevance to the test of association in any case. An exception to this is when there are only mean differences between classes (i.e condition '*3*'). In this case, for unselected samples, there is a reduction in LRT under the misclassification condition. For selected samples (scheme A) the impact of stratification under condition '*3*' seems less consistent, presumably reflecting variation in subpopulation means after selection.

Table 9.17 gives only the main results for conditions *4* and *5* (i.e. true stratification) when either there is no QTL effect ('*a*') or when there is an additive QTL effect in both classes ('*c*'). Results are similar when dominance effects and/or QTL × class interactions are simulated. The first two columns labelled "None" represent the

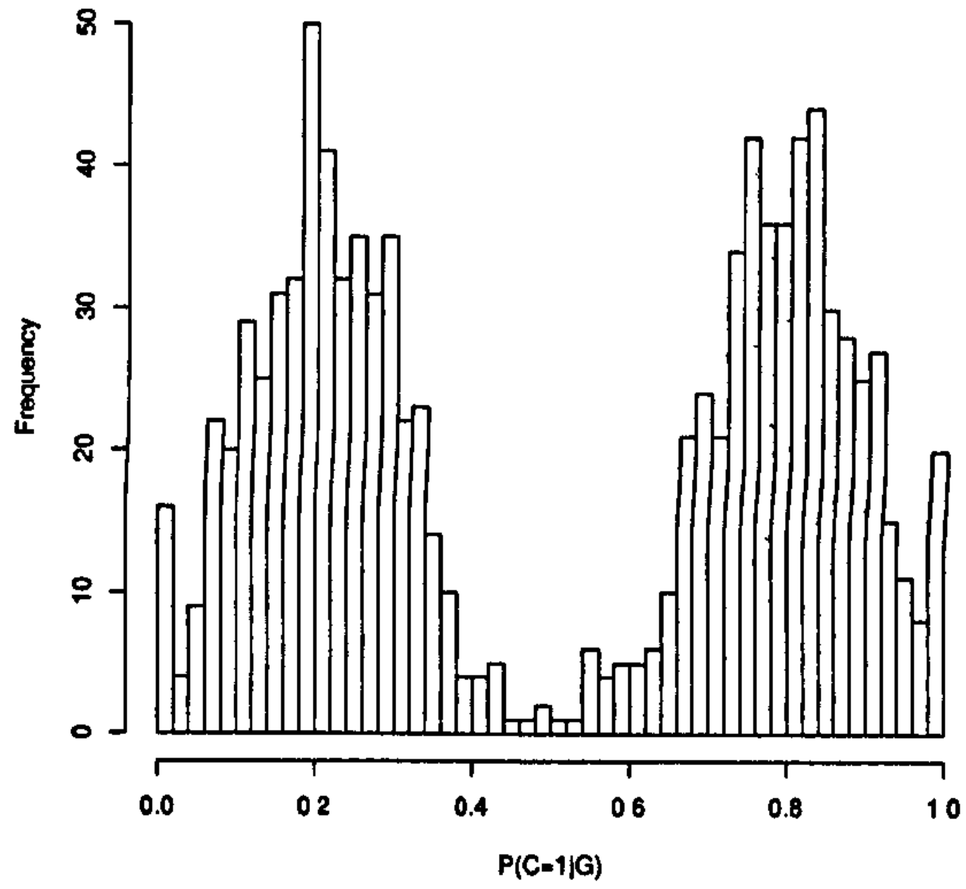


Figure 9.4: Imperfect classification of strata: example distribution of $P(C = 1|G)$ from the third 'imperfect' scheme.

	None		Perfect		Imperfect1		Imperfect2		Imperfect3	
	ML	Reg	ML	Reg	ML	Reg	ML	Reg	ML	Reg
	Full sample									
4a	39.92	40.79	2.08	2.07	2.89	2.07	21.68	22.03	3.84	2.77
4c	2.58	2.59	114.32	123.19	89.22	123.19	16.66	16.83	77.39	93.09
5a	39.51	40.38	2.14	2.12	2.81	2.12	21.67	22.04	3.78	2.66
5c	194.39	216.14	113.28	121.94	117.94	121.94	166.36	183.17	123.60	130.45
	Scheme A									
4a	15.46	16.02	2.23	1.95	2.72	1.95	11.01	11.27	3.74	2.81
4c	11.50	12.26	13.00	13.17	13.38	13.17	10.51	10.97	11.46	11.01
5a	15.97	16.60	2.20	1.99	2.59	1.99	11.62	11.89	3.76	2.73
5c	113.91	154.56	12.54	15.17	31.64	15.17	78.90	103.63	34.33	27.48

Table 9.17: The impact of imperfect classification on ML and regression approaches.

test of association that ignores strata and so is not robust. The "Perfect" columns represent the robust test using the true stratum membership values. The next set of columns "Imperfect1" to "Imperfect3" represent the results using the robust tests but with the 'imperfect' posterior probabilities, i.e. the three conditions as described above.

Taking first the ML results, the "None" and "Perfect" conditions show similar patterns of results as before. "Imperfect1" introduces an increase in the average χ^2 test statistic from 2 to around 2.8. The effects of misclassification, i.e. "Imperfect2",

are even more dramatic – the overall test statistics are very liberal when there is no QTL effect (4*a* and 5*a*). Of course, "Imperfect2" is really quite extreme – 20% of individuals are *mis*classified with complete 'certainty' which is unlikely to happen in practice. The impact of the "Imperfect3" is more similar to "Imperfect1". The test statistics are still almost doubled under the null however.

A similar pattern of results is seen for the regression method. In general, the regression method still seems to be more powerful under the alternate hypothesis and more robust under the null. Note that there is no difference between "Perfect" and "Imperfect1" for the regression method. Although this pattern of posterior probabilities is very unlikely to occur in practice (i.e. all individuals either have posterior probabilities of exactly 0.8 or 0.2) this reflects the earlier observation that, technically, the use of standard regression is inappropriate here. That is, there is no way the basic regression approach can take account of the uncertainty in stratum membership. Rather, the difference between [0,1] and [0.2,0.8] becomes only one of scaling rather than of information. A more appropriate method might involve using mixtures of regressions. However, at least in the current circumstances, the basic regression method seems to perform very well, both in unselected and selected samples.

9.4 Polygenic selection effect: 'spurious stratification'

The genetic background approach of using unlinked markers to detect and correct for population stratification, considered in Chapter 6, was designed in the context of an unselected sample. This section explores the possible impact of a specific bias that might operate when applying such methods to samples selected for extreme trait values. As samples are typically selected for extreme trait values prior to any genotyping is performed (i.e. in order to minimise costs and maximise efficiency) then

selection will always tend to precede stratification analysis.

Consider a polygenic trait in a homogeneous population, for which all contributing trait loci are both unlinked and in linkage equilibrium. In an unselected sample, these trait loci could be safely included amongst the background marker loci in the stratification analysis. Indeed, the researcher would typically be ignorant as to whether or not any one marker locus used in the stratification analysis were a trait loci or not. In many practical scenarios, e.g. using a set of genome scan markers or a panel of candidate loci, it is quite feasible that some of the markers in the stratification analysis will show an association with the trait.

Consider now that the sample has first been selected on the basis of extreme trait values prior to stratification and association analysis. The selection procedure will induce heterogeneity within the sample, in that the high "group" will have higher allele frequencies for trait-increasing alleles at all trait loci. In this way, the trait loci will be correlated within the entire selected sample. That is, increaser alleles at the different loci are more likely to cluster together (i.e. in the selected high scoring individuals) and decreaser alleles will cluster together (i.e. in the selected low scoring individuals). This could potentially be detected as a signature of stratification (i.e. unlinked loci showing linkage disequilibrium) and may therefore generate evidence for population substructure within the homogeneous sample. Furthermore, the detected substructure will be associated with the trait. If the association analysis is conducted conditional on this substructure, one would expect a drop in power, therefore.

The following simulations explore this possibility and attempt to quantify any reduction in power. A diallelic test QTL L_T is simulated with $a = 0.5$, $d = 0$ and $p = 0.5$. In addition, either 2, 5 or 10 other polygenes, L_P, are simulated such that they jointly accounted for 5%, 25%, 50% or 75% of the total trait variance. The allele frequency of L_P is either set at $P_P = 0.5$ or $P_P = 0.1$. The number of polygenes is denoted N_P, the additive genetic effect of the polygenes is denoted a_P. The remaining

N_P	P_P	σ^2_P/σ^2_T	a_P	σ^2_R	LRT	σ^2_L/σ^2_T
2	0.5	0.050	0.5500	5.5725	21.10	0.021
2	0.5	0.250	1.2250	4.3744	19.83	0.020
2	0.5	0.500	1.7325	2.8734	21.94	0.022
2	0.5	0.750	2.1220	1.3721	21.96	0.022
2	0.1	0.050	0.9100	5.5769	20.58	0.020
2	0.1	0.250	2.0400	4.3768	20.25	0.020
2	0.1	0.500	2.8860	2.8766	23.02	0.023
2	0.1	0.750	3.5350	1.3764	22.51	0.022
5	0.5	0.050	0.3450	5.5774	20.85	0.021
5	0.5	0.250	0.7750	4.3734	21.72	0.021
5	0.5	0.500	1.0950	2.8774	24.08	0.024
5	0.5	0.750	1.3420	1.3726	21.84	0.022
5	0.1	0.050	0.5800	5.5722	20.83	0.021
5	0.1	0.250	1.2900	4.3773	23.36	0.023
5	0.1	0.500	1.8250	2.8774	20.88	0.021
5	0.1	0.750	2.2360	1.3753	21.54	0.021
10	0.5	0.050	0.2450	5.5749	23.17	0.023
10	0.5	0.250	0.5475	4.3762	21.13	0.021
10	0.5	0.500	0.7745	2.8757	22.52	0.022
10	0.5	0.750	0.9485	1.3767	24.77	0.024
10	0.1	0.050	0.4100	5.5724	20.88	0.021
10	0.1	0.250	0.9125	4.3762	21.85	0.022
10	0.1	0.500	1.2910	2.8750	24.73	0.024
10	0.1	0.750	1.5815	1.3729	24.16	0.024

Table 9.18: Polygenic selection effect: values of a_P and σ^2_R used to simulate the data; also, the entire unselected sample test statistic and proportion of variance attributable to the QTL.

residual variance is set in such a way that L_T always accounted for a fixed proportion of trait variance ($\sigma^2_T \approx 0.021$). Fifty null marker loci, L_N, are also generated, with $p = 0.5$ and $a = d = 0$. The three selection schemes, A, B and C are used to select 200 individuals from the initial 1000. The stratification analyses, performed on the selected samples, use the marker set $L_P + L_N$. The solution to the stratification analysis is then used in a test of association with L_T.

Table 9.18 shows the parameter values of a_P and σ^2_R used to simulate the marker loci and trait scores under the different conditions, such that L_T explains a constant

proportion of variance. Also shown is the ML test likelihood ratio test statistic for the full unselected sample – around 22 in all cases, which is highly significant. No stratification is modelled in this case – the stratification analysis occurs after sample selection.

Table 9.19 shows the average AIC difference between a one- and two-class solution from L-POP on $L_P + L_N$ in the selected samples. A negative difference favours the one-class solution; a positive difference favours a two-class solution. In parentheses the proportion of times the two-class solution was favoured is given. For scheme A, L_P must account for at least 50% of the sample variance before a two class solution is favoured (unless $N_P = 2$ and the allele frequency is rare, 0.1). In these cases, a two-class solution is favoured almost 100% of the time. The results for schemes B and C are similar.

Figure 9.5 illustrates this effect, plotting the posterior probability of belonging to class 1 against the trait score. In this case, $N_P = 5$, $P_P = 0.5$ and $\sigma^2_P/\sigma^2_T = 0.5$, and the selection scheme is A. The majority of individuals in the low tail are unlikely to belong to class 1; in contrast, most of the high scorers are likely to belong to class 1. The sample was simulated as a homogeneous sample, however, without any stratification effects at all.

Table 9.20 gives the main results of this section. For the full sample and the selected samples the statistical power of the test of association is given. Two tests are reported for each selected sample scheme: either not controlling (pa/p) or controlling (PA/P) for strata differences. In this case, the "strata" will actually be spurious, induced by the combination of sample selection and including polygenes for the trait in the stratification analysis.

Comparing the power of the full sample against pa/p for the selected samples shows the reduction in power arising from analysing only 20% of the sample. Comparing against the PA/P model for the selected sample shows the further deterioration in

			$AIC_1 - AIC_2$		
N_P	P_P	σ^2_P/σ^2_T	Scheme A	Scheme B	Scheme C
2	0.5	0.050	-11.25 (0.08)	-8.93 (0.14)	-11.30 (0.06)
2	0.5	0.250	-4.53 (0.20)	-7.67 (0.22)	-7.91 (0.20)
2	0.5	0.500	45.02 (1.00)	19.58 (0.80)	22.31 (0.90)
2	0.5	0.750	176.38 (1.00)	129.20 (1.00)	107.30 (1.00)
2	0.1	0.050	-11.15 (0.10)	-11.48 (0.10)	-11.96 (0.14)
2	0.1	0.250	-6.05 (0.32)	-9.62 (0.10)	-8.61 (0.12)
2	0.1	0.500	-9.60 (0.12)	-11.40 (0.04)	-5.73 (0.28)
2	0.1	0.750	-8.56 (0.18)	-9.80 (0.10)	-6.14 (0.28)
5	0.5	0.050	-8.65 (0.26)	-11.80 (0.08)	-9.78 (0.18)
5	0.5	0.250	-2.53 (0.34)	-6.94 (0.24)	-8.32 (0.22)
5	0.5	0.500	37.08 (1.00)	16.47 (0.84)	23.32 (0.86)
5	0.5	0.750	162.53 (1.00)	110.55 (1.00)	94.90 (1.00)
5	0.1	0.050	-10.39 (0.16)	-9.61 (0.14)	-13.49 (0.04)
5	0.1	0.250	-9.04 (0.18)	-10.90 (0.10)	-8.94 (0.20)
5	0.1	0.500	9.15 (0.60)	-10.06 (0.16)	-0.43 (0.48)
5	0.1	0.750	74.26 (0.98)	-11.48 (0.06)	32.12 (0.92)
10	0.5	0.050	-10.01 (0.12)	-9.85 (0.14)	-10.72 (0.14)
10	0.5	0.250	-5.27 (0.26)	-7.74 (0.18)	-8.73 (0.32)
10	0.5	0.500	42.04 (0.96)	14.94 (0.74)	20.20 (0.90)
10	0.5	0.750	144.53 (1.00)	97.22 (1.00)	81.93 (1.00)
10	0.1	0.050	-10.94 (0.16)	-11.62 (0.14)	-12.77 (0.10)
10	0.1	0.250	-6.37 (0.18)	-11.71 (0.12)	-7.12 (0.16)
10	0.1	0.500	27.13 (0.90)	-8.02 (0.26)	4.55 (0.48)
10	0.1	0.750	136.62 (1.00)	13.25 (0.47)	75.97 (1.00)

Table 9.19: L-POP results on the selected samples: difference in AIC between a $K = 1$ and a $K = 2$ solution, in parentheses the proportion of times a $K = 2$ solution is favoured.

power due to modelling the spurious stratification. As expected, the cases when there is a further deterioration due to spurious stratification are largely the same cases as when L-POP consistently extracts a two-class solution – roughly speaking when the polygenes account for at least 50% of the trait variance. For example, if $N_P = 5$, $P_P = 0.5$ and $\sigma^2_P/\sigma^2_T = 0.5$ then power drops from 97% in the full sample to 86% in the selected sample A when stratification is not modelled. When stratification is modelled, power drops further to 72%. When the polygenes account for 75% of the trait variance, power can drop to as low as 50% in selected samples. The results are

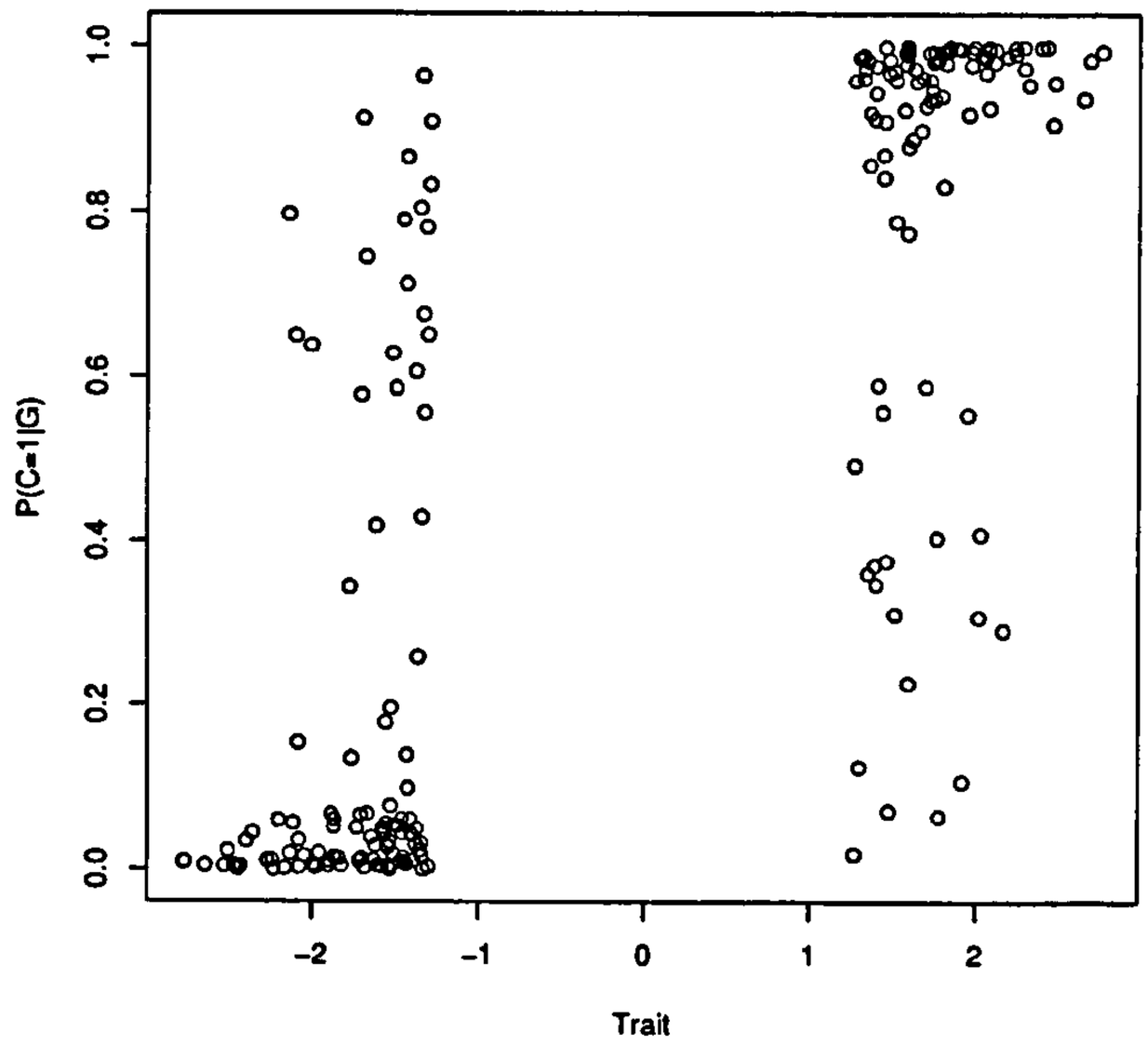


Figure 9.5: Population stratification in selected samples: a plot of the posterior class probabilities by trait score, where sample selection has induced a 'spurious stratification'.

based on all replicates – i.e. not just those where L-POP extracted a two-class solution. However, the pattern of results stays much the same when stratified by whether or not a one- or a two-class solution was favoured.

In summary, this section has demonstrated the possibility of a spurious stratification effect arising from performing genetic background analyses on polygenic traits in selected samples. Whether this phenomenon is likely to occur in practice is another issue. Selecting specific markers for stratification analysis (that are both most likely to be functionless and show the greatest frequency differences between ethnic groups) is likely to be the best course of action.

Of course, it is not desirable to remove from the stratification analysis marker set all markers that show an association with the trait. To do so would render the stratification analysis useless, as it precisely relies on multiple markers that will show an association with the trait if there is a mean difference between strata.

			Full	Scheme A		Scheme B		Scheme C	
N_P	P_P	σ^2_P/σ^2_T	pa/p	pa/p	PA/P	pa/p	PA/P	pa/p	PA/P
2	0.5	0.050	0.98	0.92	0.90	0.85	0.82	0.88	0.90
2	0.5	0.250	0.99	0.93	0.89	0.89	0.87	0.87	0.89
2	0.5	0.500	0.95	0.91	0.74	0.82	0.73	0.86	0.83
2	0.5	0.750	0.98	0.93	0.53	0.91	0.59	0.87	0.80
2	0.1	0.050	0.96	0.89	0.87	0.82	0.79	0.90	0.92
2	0.1	0.250	0.98	0.91	0.90	0.89	0.88	0.85	0.87
2	0.1	0.500	0.97	0.90	0.87	0.86	0.83	0.90	0.92
2	0.1	0.750	0.97	0.96	0.95	0.94	0.93	0.95	0.95
5	0.5	0.050	0.97	0.85	0.82	0.76	0.74	0.86	0.88
5	0.5	0.250	0.96	0.91	0.88	0.84	0.81	0.91	0.92
5	0.5	0.500	0.97	0.86	0.72	0.78	0.67	0.88	0.83
5	0.5	0.750	0.98	0.85	0.52	0.80	0.53	0.85	0.69
5	0.1	0.050	0.97	0.87	0.84	0.80	0.80	0.82	0.85
5	0.1	0.250	0.96	0.86	0.83	0.84	0.81	0.86	0.88
5	0.1	0.500	0.96	0.92	0.79	0.85	0.85	0.88	0.89
5	0.1	0.750	0.97	0.93	0.53	0.91	0.89	0.90	0.79
10	0.5	0.050	0.96	0.89	0.88	0.83	0.82	0.83	0.86
10	0.5	0.250	0.97	0.90	0.88	0.82	0.79	0.83	0.86
10	0.5	0.500	0.97	0.91	0.78	0.85	0.78	0.87	0.85
10	0.5	0.750	0.97	0.87	0.56	0.80	0.46	0.81	0.78
10	0.1	0.050	0.98	0.88	0.88	0.87	0.85	0.86	0.88
10	0.1	0.250	0.96	0.89	0.87	0.80	0.79	0.87	0.87
10	0.1	0.500	0.98	0.90	0.82	0.86	0.86	0.87	0.88
10	0.1	0.750	0.97	0.93	0.61	0.90	0.79	0.89	0.85

Table 9.20: Polygenic selection effects: power of the pa/p and PA/P ML association tests.

9.5 Summary

The simulation studies in the Chapter have illustrated two methods of testing for association in selected samples which can correct for effects of populations stratification. The standard regression method performed marginally better than the ML method in most circumstances. An exception to this is when the data are from non-normal distributions (whether or not the sample is a selected one). In this case, modelling genotype conditional on trait score gives a test with greater power. To relate these results to Chapter 3, the present regression test is only robust because it is for total

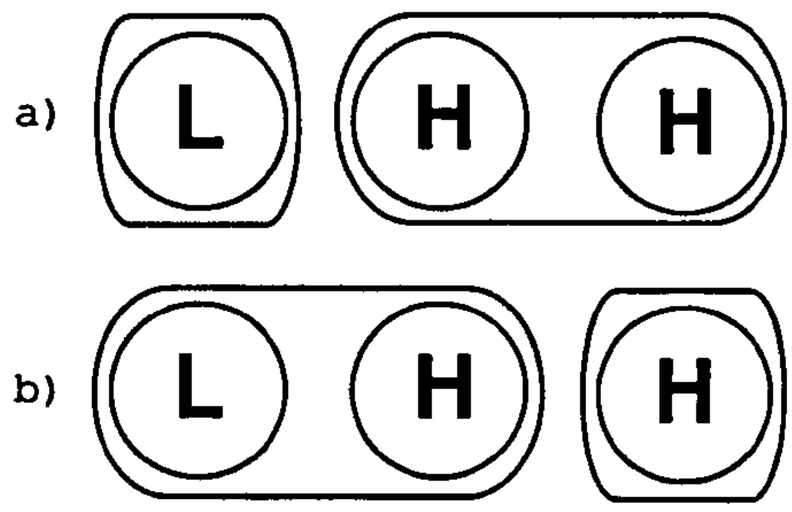


Figure 9.6: Discrepancies between true and estimated population substructure.

association in unrelated individuals. As demonstrated in that Chapter, if the test were for the robust within-family component of association, then the standard regression approach would not be valid.

Two particular scenarios were highlighted in which the benefit of using selected samples is reduced: first, that controlling for stratification effects can lead to a reduction in power to detect true QTL in selected samples. Second, that performing stratification on selected samples can induce a 'spurious stratification' if polygenes for the trait are included in the 'null' marker set.

With imperfect classification the robustness of both ML and regression methods deteriorates. Furthermore, it is conceivable that the pattern of population substructure detected by the stratification analysis may only partially overlap with the strata effects concerning the test locus. For example, Figure 9.6 illustrates two scenarios in which the wrong solution has been selected by stratification analysis (i.e. a $K = 2$ solution instead of a $K = 3$ solution). In reality, there are three subpopulations (white circles): a "H" represents a high trait mean and a high allele frequency at the test locus. Likewise, a "L" represents a low trait mean and a low allele frequency. Otherwise, the three subpopulations may or may not be particularly genetically distinct: in both scenarios, a $K = 2$ solution has been selected (estimated classes represented by the gray ovals). There is no true association between trait and the test locus.

If two subpopulations are not particularly distinct they may well be pooled into the same estimated class in stratification analysis. In the first case, a), this would

not present a problem for the association analysis – if one were to control for the two estimated classes, the nature of the stratification at the test locus would still be captured. If, however, scenario b) were true, then even if one controlled for estimated class structure, there could still be false positives due to population stratification effects within estimated class. It seems possible that two subpopulations could be genetically very similar across most of the genome but differ markedly for a particular trait and particular test locus.

In this sense, genetic background methods do not provide the 'logical' protection from population stratification that family-based methods do. Is there any utility in the genetic background approach to population stratification then? If enough markers with the right properties are used for the stratification analyses, such that one could assume all the significant substructure in the sample has been captured, then there do seem to be advantages to this approach. "Enough markers" appears to be at least 100, based on simulation results reported in Chapter 6, although this figure can be dramatically reduced if the "right" markers are chosen: preferably multi-allelic markers that are known to show large allele frequency differences between ethnic groups. In this case, the advantages are a more efficient design that is easier to collect (i.e. unrelated individuals versus families) and, secondly, the ability to allow for stratum-specific QTL effects. Allowing for such effects should increase power to detect QTL; also, such findings would be of considerable interest in themselves. Such QTL × strata interactions are not detectable within standard family-based association methodologies.

Chapter 10

Conclusion

This thesis has examined several areas of quantitative trait locus analysis, with an emphasis on the utility of selected samples and the complex nature of multifactorial human traits. A strategy for the selection of sibships for linkage and association analyses was presented in Chapters 2 and 3, as well as methods to analyse selected samples. The complex effects of gene–environment interaction, epistasis and population stratification were examined, in various contexts, in Chapters 4, 5 and 6. The final three Chapters re-examined these three complex effects with a special emphasis on selected samples.

In these concluding remarks, I would like to briefly consider the utility of *unselected* samples. Broadly speaking, sample selection offers a substantial gain in efficiency for the detection of simple, main effects. However, unselected samples potentially afford a number of other advantages. Firstly, it is important to remember that selecting on the basis of trait scores is only beneficial when the relative cost of phenotyping versus genotyping is low. If the phenotype is based on a self-report postal questionnaire, this may be the case; if it is based on an MRI scan, this would not be so (although it would of course be possible to select on related phenotypes prior to the MRI scan).

There are a number of potential problems related to the use of selected samples. For example, the use of discordant sib pairs will enrich a sample for instances of non-

paternity, as half-siblings will be genotypically and phenotypically more dissimilar than full siblings. Given the already high non-paternity rates in most populations, this could potentially be a significant problem. Although, in the context of a genome scan, it is relatively easy to detect half-siblings and analyse them appropriately, half-siblings, even if phenotypically discordant, are likely to offer less information than a randomly-selected full sibling pair.

There are also issues with how 'extreme' extreme selection should be. As Lander and Botstein (1989) noted, individuals in the extreme tails of a distribution are perhaps more likely to represent instances of measurement error or other artefactual cases. Allison et al. (1998) indicate another situation in which more extreme sampling does not necessarily result in more powerful samples. By use of simulation, they show that 'extremely extreme' sampling can reduce power for both linkage and association under certain (equally extreme) oligogenic models. Essentially, Allison et al. (1998) note that the presence of a major, non-additive effect at a second locus can induce non-normality in the residual, within-genotype distributions of the first locus. This, they argue, can lead to a reduction in power (although the conditioning-on-trait approach described in this thesis should not be so susceptible to such effects). However, the conditions they simulate involve particularly extreme genetic models: for example, a second locus with a rare allele that has a displacement effect of 4 standard deviations. Figure 10.1 schematically illustrates the effect of a second major effect (dominant locus B) on the power to detect the additive QTL A. In the left panel, without locus B, sampling individuals above the threshold (the vertical line) would enrich for the a allele, and (within the context of a complete sampling strategy, e.g. also sampling low individuals) this would increase power. In the right panel, the effect of locus B effectively masks that of locus A – in this case, we would expect to observe only the population frequencies for locus A in the selected sample.

In the kind of scenario considered by Allison et al. (1998), the major locus needed

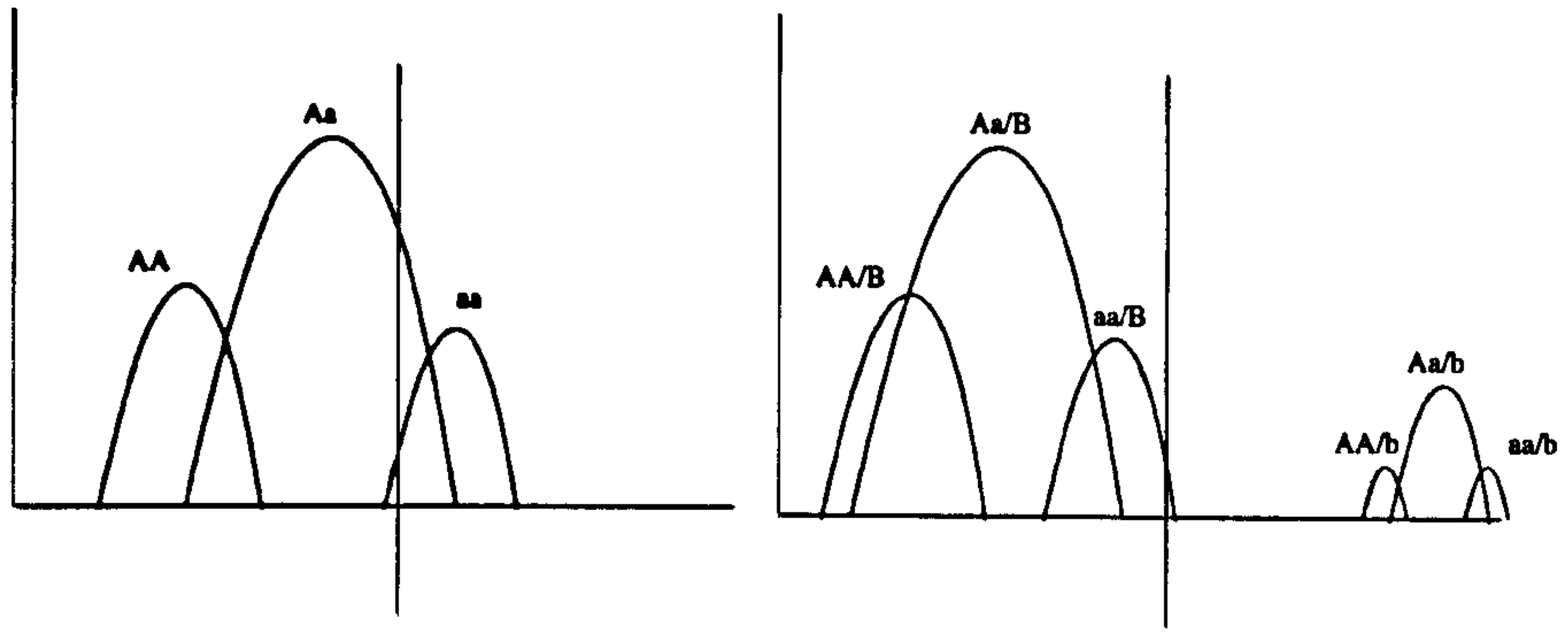


Figure 10.1: Extreme sample selection under oligogenic models. In the left figure, extreme sample selection, i.e. selecting individuals above the threshold (vertical line) would result in greater power to detect locus A. In the right panel, which also includes a major dominant effect of locus B, power to detect locus A would not increase as a consequence of selection.

to induce this kind of effect would presumably be easily detectable by standard analytic methods. Any subsequent analysis (and selection) could be performed conditional on this locus. More generally, however, this points to the need to adjust for as many relevant factors as possible prior to selection. If, for example, there is a marked sex difference for a quantitative trait, then selecting high and low groups may simply be enriching the selected groups for males and females respectively, rather than particular trait-influencing alleles.

Selecting on extremes might also mean that genes controlling normal variation will go undetected, and only genes for extreme phenotypes will be found. Allison et al. (1998) give the example of height: extreme sample selection might find rare genes for Marfan syndrome or achondroplastic dwarfism, but not the common variants that control height in the normal population. An unselected sample would potentially allow the investigator to test whether a given QTL operates systematically across the entire trait or only at extremes. These issues underscore the important value of appropriate phenotypic definitions. For example, whether or not the extreme values of a continuum represent a qualitatively different subtype or not is a phenotypic question that would ideally be answered prior to selecting for linkage and association studies. Twin and family studies can be useful in addressing such questions, (e.g. DeFries-

Fulker extremes analysis: DeFries and Fulker, 1985, 1988; Purcell and Sham, 2002). Of course, even in unselected samples, extreme individuals will be more influential in analysis, and so these issues apply to both unselected and selected designs.

As mentioned in Chapter 3, there are also further considerations relating to the effects of sample selection on linkage disequilibrium mapping (Abecasis et al., 2001a). That is, allele frequencies at both the marker and the QTL have implications for the power of LD mapping, and sample selection can potentially impact on these. Using balanced, symmetrical designs should minimise any such effects, however.

The above issues relate to the detection of univariate, additive QTL effects. Further issues with selected samples arise when considering multivariate applications. In large studies, there will most likely be more than one phenotype of interest. Adopting a selected sample approach typically forces the investigator to look at only a single phenotype, or a cluster of strongly-related measures. Other measures could potentially be included as dependent variables in separate analyses, combined in multivariate analyses, or used as modifier variables in interaction analyses. Obviously, if the other measures are not highly correlated with the trait used to select the sample, then attempts to map QTL for these other measures will not benefit from increased efficiency due to the selection. Indeed, the selection on one variable might even select for atypical cases on the second variable, e.g. the population of individuals who are depressed *and* heavy drinkers may not be representative of the population of those who are just heavy drinkers. Alternatively, if the other measures are correlated with the main trait, the potential to look for interaction effects in selected samples might be severely limited. That is, the required variation in the potential moderator variables might have vanished in the selected sample, e.g. as height and sex are associated, there may be no very tall women or very short men in a sample selected for extreme height, which would not help height × sex interaction analyses.

It is of course wrong to advocate any single experimental design as 'optimal'.

Studies which solely aim to detect QTL will be better served by different designs than studies that attempt to more comprehensively 'dissect the genetic architecture' of complex traits. Although the majority of research areas are still at the first stage, the next decade will undoubtedly see studies moving towards the second goal, of piecing together individual genetic and environmental risk factors into coherent models for human traits and diseases. One particularly promising direction of research is embodied in the UK's BioBank project, which aims specifically to detect genetic and environmental interactions for a multitude of complex diseases: the study design is based on sampling 500,000 *unselected* individuals. With complete phenotypic and genotypic data on this number of individuals, many of the hard statistical issues we currently face (e.g. dealing with ascertainment, developing optimally powerful tests) may be of lesser importance. In the mean time, methods are still needed to fully exploit the moderately-sized, incomplete samples that exist today.

Bibliography

G. R. Abecasis, L. R. Cardon, and W. O. Cookson. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66: 279 – 292, 2000.

G. R. Abecasis, W. O. C. Cookson, and L. R. Cardon. Selection strategies for disequilibrium mapping. *American Journal of Human Genetics*, 68:1463 – 1474, 2001a.

G. R. Abecasis, E. Noguchi, A. Heinzmann, J. A. Traherne, S. Bhattacharyya, N. I. Leaves, G. G. Anderson, Zhang Y., N. J. Lench, A. Carey, L.R. Cardon, M.F. Moffatt, and W.O. Cookson. Extent and distribution of linkage disequilibrium in three genomic regions. *American Journal of Human Genetics*, 69(1):191 – 197, 2001b.

H. Akaike. A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, AC-19:716 – 723, 1974.

A. Alcais and L. Abel. Maximum-likelihood binomial method of genetic model-free linkage analysis of quantitative traits in sibships. *Genetic Epidemiology*, 17:102 – 117, 1999.

A. Alcais and L. Abel. Linkage analysis of quantitative trait loci: sib pairs or sibships? *Human Heredity*, 50:251 – 256, 2000.

D. B. Allison, M. Heo, N. Kaplan, and E. R. Martin. Sibling-based tests of linkage

and association for quantitative traits. *American Journal of Human Genetics*, 64: 1754 – 1764, 1999a.

D. B. Allison, M. Heo, N. J. Schork, S-L. Wong, and R. C. Elston. Extreme selection strategies in gene mapping studies of oligogenic quantitative traits do not always increase power. *Human Heredity*, 48:97 – 107, 1998.

D. B. Allison, M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos, and J. Blangero. Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. *American Journal of Human Genetics*, 65:531 – 544, 1999b.

L. Almasy and J. Blangero. Multipoint quantitative-trait linkage analysis in general pedigrees. *American Journal of Human Genetics*, 62:1198–1211, 1998.

C. I. Amos. Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics*, 54:535–543, 1994.

C. J. Amos and M. de Andrade. Genetic linkage methods for quantitative traits. *Statistical Methods in Medical Research*, 10:3 – 25, 2001.

S.A. Bacanu, B. Devlin, and K. Roeder. The power of genomic control. *American Journal of Human Genetics*, 66:1933 – 1944, 2000.

J.S. Bader, A. Bansal, and P. Sham. Efficient SNP-based tests of association for quantitative phonotypes using pooled DNA. *Genescreen*, 2002. in press.

D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and it implications for investigating identity and paternity. *Genetica*, 96:3–12, 1995.

J. Blangero, J. T. Williams, and L. Almasy. Variance components methods for detecting complex trait loci. *Advances in Genetics*, 42:151–181, 2000.

D. I. Boomsma. Using multivariate genetic modeling to detect pleiotropic quantitative trait loci. *Behavior Genetics*, 26:161–166, 1996.

D. I. Boomsma, E. J. C. de Geus, G. C. M. van Baal, and J. R. Koopmans. A religious upbringing reduces the influence of genetic factors on disinhibition: evidence for interaction between genotype and environment on personality. *Twin Research*, 2 (2):115 – 125, 1999.

A. M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, and L.L. Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368:455 – 457, 1994.

L.R. Cardon and J.I. Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91 – 99, 2001.

G. Carey and J.A. Williamson. Linkage analysis of quantitative traits: increased power by using selected samples. *American Journal of Human Genetics*, 49:786 – 796, 1991.

M. Cargil, D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, and et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22:239 – 247, 1999.

L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. *The History and Geography of Human Genes.* Princeton Univ. Press, Princeton, NJ., 1994.

L.L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: models and estimation procedures. *Evolution*, 21:550–570, 1967.

K.H. Cheung, P.L. Miller, J.R. Kidd, K.K. Kidd, M.V. Osier, and A.J. Pakstis. ALFRED: a Web-accessible allele frequency database. *Pac Symp Biocomput*, pages 639 – 650, 2000.

C. C. Cockerham. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39:859 – 882, 1954.

M. J. Crowder and D. J. Hand. *Analysis of Repeated Measures.* Chapman & Hall, 1990.

R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich. A Perspective on Epistasis: Limits of Models Displaying No Main Effect. *American Journal of Human Genetics*, 70: 461 – 471, 2002.

M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nature Genetics*, 29:229 – 232, 2001.

J. Daniels, P. Holmans, N. Williams, and et al. A simple method for analysing microsatellite allele image patterns generated from dna pools and its applications to allelic association studies. *American Journal of Human Genetics*, 62:1189 – 1197, 1998.

M. Dean, J. C. Stephens, C. Winkler, D. A. Lomb, M. Ramsburg, R. Boaze, C. Stewart, L. Charbonneau, D. Goldman, B. J. Albaugh, and et al. Polymorphic admixture typing in human ethnic populations. *American Journal of Human Genetics*, 55:788 – 808, 1994.

J. C. DeFries and D. W. Fulker. Multiple regression analysis of twin data. *Behavior Genetics*, 15(5):467 – 473, 1985.

J. C. DeFries and D. W. Fulker. Multiple regression analysis of twin data: etiology of deviant scores versus individual differences. *Acta Genet Med Gemellol*, 37:205 – 216, 1988.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistial Society*, B 39(1):1–38, 1977.

B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55: 997 – 1004, 1999.

C. V. Dolan and D. I. Boomsma. Optimal selection of sib pairs from random samples for linkage analysis of a QTL using the EDAC test. *Behavior Genetics*, 28(3): 197–206, 1998.

C.V. Dolan, D.I. Boomsma, and M.C. Neale. A simulation study of the effects of assignment of prior identity-by-descent probabilities to unselected sib pairs, in covariance-structure modeling of a quantitative-trait locus. *American Journal of Human Genetics*, 64:268 – 280, 1999.

S. Dudoit and T. P. Speed. A score test for linkage using identity by descent data from sibships. *Annals of Statistics*, 27:943 – 986., 1999.

S. Dudoit and T. P. Speed. A score test for linkage analysis of qualitative and quantitative traits based on identity by descent on sib-pairs. *Biostatistics*, 1:1–26, 2000.

G. Dunn, B. Everitt, and A. Pickles, editors. *Modelling covariances and latent variables using EQS.* Chapman Hall, London., 1993.

L. Eaves and A. Erkanli. Markov Chain Monte Carlo approaches to analysis of genetic and environmental components of human developmental change and GxE interaction. *Behavior Genetics*, (in press), 2002.

L. Eaves and J. Meyer. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics*, 24(5):443 – 455, 1994.

L. J. Eaves. Effect of genetic architecture on the power of human linkage studies to resolve the contribution of quantitative trait loci. *Heredity*, 72:175 – 192, 1994.

L. J. Eaves, K. Last, N. G. Martin, and J. L. Jinks. A progressive approach to non-addivity and genotype-environmental covariance in the analysis of human differences. *British Journal of Mathematical and Statistical Psychology*, 30:1 – 42, 1977.

R. C. Elston. Linkage and association. *Genetic Epidemiology*, 15:565 – 576, 1998.

R. C. Elston and E. Sobel. Sampling considerations in the gathering and analysis of pedigree data. *American Journal of Human Genetics*, 31:62 – 69, 1979.

R. C. Elston and J. Stewart. A general model for the analysis of pedigree data. *Human Heredity*, 21:523 – 542, 1971.

D. Evans. The power of multivariate quantitative-trait loci linkage analysis is influenced by the correlation between variables. *American Journal of Human Genetics*, 70:1599 – 1602, 2002.

W. J. Ewens and N. C. E. Shute. A resolution of the ascertainment sampling problem. *Theoretical Population Biology*, 30:388 – 412, 1986.

D. S. Falconer. *Introduction to Quantitative Genetics*. Longman Scientific & Technical, London, 3rd edition, 1989.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, 222:309 – 368, 1922.

R. A. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

S. A. Fisher and C. M. Lewis. Methods to identify population outliers using genetic markers. *GeneScreen*, 1:125 – 129, 2001.

W. N. Frankel and N. J. Schork. Who's afraid of epistasis? *Nature Genetics*, 14:371 – 373, 1996.

D. W. Fulker and S. S. Cherny. An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics*, 26:527–532, 1996.

D.W. Fulker, S.S. Cherney, P.C. Sham, and J.K. Hewitt. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*, 64(1):259 – 267, 1999.

W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex bayesian modelling. *The Statistician*, 43:169 – 178, 1994.

C. Gu, A. Todorov, and D.C. Rao. Combining extremely concordant sibpairs with extremely discordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci. *Genetic Epidemiology*, 13:513 – 533, 1996.

R. Guerra, Y. Wan, A. Jia, C. I. Amos, and J. C. Cohen. Linkage testing under robust genetic models. *Human Heredity*, 49:146 – 153, 1999.

M. C. Gurganus, S. V. Nuzhdin, J. W. Leips, and T. F. C. Mackay. High-resolution mapping of quantitative trait loci for sternopleural bristle number in Drosophila melanogaster. *Genetics*, 152:1585 – 1604, 1999.

J. B. S. Haldance. The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, 8:229 – 309, 1919.

M. K. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, 22:239 – 247, 1999.

J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3–19, 1972.

A. C. Heath, L. J. Eaves, and N. G. Martin. Interaction of marital status and genetic risk for symptoms of depression. *Twin Research*, 1:119 – 122, 1998.

A. C. Heath, A. A. Todorov, E. C. Nelson, P. A. F. Madden, K. K. Bucholz, and N. G. Martin. Gene-environment interaction effects on behavioral variation and risk of complex disorders: The example of alcoholism and other psychiatric disorders. *Twin Research*, 5(1):30 – 37, 2002.

S. E. Hodge and R. C. Elston. Lods, wrods and mods: the interpretation of lod scores calculated under different models. *Genetic Epidemiology*, 11:329 – 342, 1994.

J. L Hopper. Variance components for statistical genetics: applications in medical research to characteristics related to human diseases and health. *Statistical Methods in Medical Research*, 2:199 – 223, 1993.

J. L. Hopper and J. D. Matthews. Extensions to multivariate normal models for pedigee analysis. *Annals of Human Genetics*, 39:485 – 491, 1982.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, pages 193 – 218, 1985.

R. M. Huggins. On the robust analysis of variance components models for pedigree data. *Austrailian Journal of Statistics*, 51:178 – 190, 1993.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860 – 921, 2001.

K. L. Jang, P. A. Vernon, W. J. Livesley, M. B. Stein, and H. Wolf. Intra- and extra-familial influences on alcohol and drug misuse: a twin study of gene-environment correlation. *Addiction*, 96:1307 – 1318, 2001.

R. C. Jansen. Complex plant traits: time for polygenic analysis. *Trends in Plant Science*, 1(3):89 – 94, 1996.

A. Jawaid, J.S. Bader, S. Purcell, S.S. Cherny, and P.C. Sham. Optimal selection strategies for QTL mapping using pooled DNA samples. *European Journal of Human Genetics*, 10(2):125 – 132, 2002.

J. L. Jinks and D. W. Fulker. Comparison of the biometrical genetical, MAVA, and classical approaches to the analysis of behavior. *Psychol Bull*, 73:311–349, 1970.

L.B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Research*, 10(10):1435 – 1444, 2000.

K. S. Kendler and L. J. Eaves. Models for the joint effect of genotype and environment on liability to psychiatric illness. *American Journal of Psychiatry*, 143:279 – 289, 1986.

W. C. Knowler, R. C. Williams, D. J. Pettitt, and A. G. Steinberg. Gm3-5,13,14 and Type 2 diabetes-mellitus – An association in American-Indians with genetic admixture. *American Journal of Human Genetics*, 43:520 – 526, 1988.

K. Kojima. Role of epistasis and overdominance in stability of equilibria with selection. *Porc Natl Acad Sci USA*, 45:984 – 989, 1959.

L. Kruglyak and E. S. Lander. Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57:439–454, 1995a.

L. Kruglyak and E.S. Lander. A nonparametric approach for mapping quantitative trait loci. *Genetics*, 139:1421 – 1428, 1995b.

E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA*, 84:2263 – 2367, 1987.

E. S. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics*, 11:241 – 247, 1995.

E.S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185 – 199, 1989.

E.S. Lander and N.J. Schork. Genetic dissection of complex trait. *Science*, 265:2037 – 2047, 1994.

K. Lange, J. Westlake, and M.A. Spence. Extensions to pedigree analysis iii: Variance components by the scoring method. *Annals of Human Genetics*, 39:485 – 491, 1976.

K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modelling using the *t* distribution. *Journal of the American Statistical Association*, 84:881 – 896, 1989.

P. F. Lazarsfeld and N. W. Henry, editors. *Latent structure analysis.* Houghton Mifflin, Boston, 1968.

J. Leips and T. F. C. Mackay. Quantitative trait loci for lifespan in *drosophila melanogaster*: interactions with genetic background and larval density. *Genetics*, 155:1773 – 1788, 2000.

C. C. Li. Population subdivision with respect to multiple alleles. *Annals of Human Genetics*, 33:23-29, 1969.

W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50:334 – 349, 2000.

M.L. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, Massachusetts, 1st edition, 1998.

T. F. C. Mackay. Quantitative trait loci in Drosophila. *Nat Rev Genet*, 2:11 – 20, 2001.

N. G. Martin, L. J. Eaves, and A. C. Heath. Prospects for detecting genotype x environment interactions in twins with breast cancer. *Acta Genet Med Gemellol*, 36:5 – 20, 1987.

K. Mather. Non-allelic interactions in continuous variation of randomly breeding populations. *Heredity*, 32:414 – 419, 1974.

K. Mather and J. L. Jinks. *Biometrical Genetics.* Chapman and Hall, London, 3rd edition, 1982.

G. Mendel. *Experiments in plant hybridisation: English Translation and Commentary by R. A. Fisher (1965).* Oliver and Boyd, Edinburgh, 1866.

B.D. Mitchell, S. Ghosh, J.L. Schneider, G. Birznieks, and J. Blangero. Power of variance component linkage analysis to detect epistasis. *Genetic Epidemiology*, 14: 1017 – 1022, 1997.

S. A. Monks, N. L. Kaplan, and B. S. Weir. A comparative study of sibship tests of linkage and/or association. *American Journal of Human Genetics*, 63(5):1507 – 1516, 1998.

N. E. Morton. Sequential tests for the detection of linkage. *American Journal of Human Genetics*, 7:277–318, 1955.

N. E. Morton and C. J. MacLean. Analysis of family resemblance III. Complex segregation analysis of quantitative traits. *American Journal of Human Genetics*, 26:489 – 503, 1974.

M. C. Neale and L. R. Cardon. *Methodology for Genetic Studies of Twins and Families.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

M. C. (1997) Neale. *Mx: Statistical modeling.* Dept of Psychiatry, Box 980126 VCU, Richmond VA 23298, 4th edition edition, 1997.

M. Nei. *Molecular Evolutionary Genetics.* Columbia University Press, New York, 1987.

J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308 – 313, 1965.

R. J. Neuman and J. P. Rice. Two-locus models of diseases. *Genetic Epidemiology*, 9:347 – 365, 1992.

J. Ott. *Analysis of Human Genetic Linkage.* Johns Hopkins Press, Baltimore, 1991.

L. Peltonen and V. McKusick. Dissecting human disease in the postgenomic era. *Science*, 291:1224 – 1229, 2001.

L. S. Penrose. The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. *Annals of Eugenics*, 6:133 – 138, 1935.

S. Petrill, R. Plomin, G. E. McClearn, D. L. Smith, S. Vignetti, M. J. Chorney, K. Chorney, L. A. Thompson, D. K. Detterman, C. Benbow, D. Lubinski, J. Daniels, M. Owen, and P. McGuffin. No association between general cognitive ability and the A1 allele of the D2 dopamine receptor gene. *Behaviour Genetics*, 27:29 – 31, 1997.

R. Plomin, J. C. DeFries, and J. C. Loehlin. Genotype-environment interaction and correlation in the analysis of human behavior. *Psychological Bulletin*, 84:309–322, 1977.

R. Plomin, J.C.DeFries, G.E. McClearn, and P. McGuffin. *Behavioral Genetics.* Worth, 4th edition, 2001.

R. Plomin, M. J. Owen, and P. McGuffin. The genetic basis of complex human behaviors. *Science*, 264:1733–1739, 1994.

J. K. Pritchard and P. Donnelly. Case-Control Studies of Association in Structured or Admixed Populations. *Theoretical Population Biology*, 60(3):227 – 237, 2001.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945 – 959, 2000.

J.K. Pritchard and N.A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics*, 65(1):220 – 228, 1999.

S. Purcell, S. S. Cherny, and P. C. Sham. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics (in press)*, 2003.

S. Purcell and P. Sham. A model-fitting implementation of the DeFries-Fulker model for selected twin data. *Behavior Genetics*, (in press), 2002.

D. C. Rao, B. J. D. Keats, N. E Morton, and et al. Variability of human linkage data. *American Journal of Human Genetics*, 30:516 – 529, 1979.

F. V. Rijsdijk, J. K. Hewitt, and P. C. Sham. Analytic power calculation for QTL linkage analysis of small pedigrees. *European Journal of Human Genetics*, 9:335 – 340, 2001.

N. Risch. Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics*, 46:222 – 228, 1990.

N. Risch, E. Burchard, E. Ziv, and H. Tang. Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*, 3(7):1 – 12, 2002.

N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516 – 1517, 1996.

N. Risch, D. Spiker, L. Lotspeich, N. Nouri, D. Hinds, J. Hallmayer, L. Kalaydjieva, and et al. A genomic screen of autism: evidence for a multilocus etiology. *American Journal of Human Genetics*, 65:493 – 507, 1999.

N. Risch and H. Zhang. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science*, 268:1584 – 1589, 1995.

N.J. Risch and H. Zhang. Mapping quantitative trait loci with extreme discordant sib pairs: sampling considerations. *American Journal of Human Genetics*, 58:836 – 843, 1996.

N. A. Rosenberg, J. K. Pritchard, J. L. Weber, H. C. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman. Genetic structure of human populations. *Science*, 298:2381 – 2385, 2003.

G.A. Satten, D. Flanders, and Q. Yang. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics*, 68:466 – 477, 2001.

S. Scarr. Developmental theories for the 1990s: development and individual differences. *Child Development*, 63:1 – 19, 1992.

N. J. Schork. Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modelling considerations. *American Journal of Human Genetics*, 53:1306 – 1319, 1993.

N. J. Schork, D. Fallin, B. Thiel, X. Xu, U. Broekel, H. J. Jacob, and D. Cohen. The future of genetic case-control studies. *Adv Genet*, 42:191 – 212, 2001.

P. C. Sham, J. S. Bader, I. Craig, M. O' Donovan, and M. Owen. DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3:862–871, 2002a.

P. C. Sham, J. H. Zhao, S.S. Cherny, and J.K. Hewitt. Variance components qtl linkage analysis: conditioning on trait values. *Genetic Epidemiology*, 19(S1):S22–S28, 2000a.

P.C. Sham, S.S. Cherny, S. Purcell, and J.K. Hewitt. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *American Journal of Human Genetics*, 66:1616 – 1630, 2000b.

P.C. Sham and S. Purcell. Equivalence between Haseman-Elston and variance components linkage analysis for sib pairs. *American Journal of Human Genetics*, 68: 1527 – 1532, 2001.

P.C. Sham, S. Purcell, S. S. Cherny, and G. R. Abecasis. Regression-Based QTL Analysis of General Pedigrees. *American Journal of Human Genetics*, 71:238 – 253, 2002b.

P.C. Sham, A. Sterne, S. Purcell, S.S. Cherny, M. Webster, F. Rijsdijk, P. Asherson, D. Ball, I. Craig, T. Eley, D. Goldberg, J. Gray, A. Mann, M. Owen, and R. Plomin. Genesis: Creating a composite index of the vulnerability to anxiety and depression in a community-based sample of siblings. *Twin Research*, 3:316 – 322, 2001.

P. A. Silva and W. Stanton, editors. *From child to adult.* Oxford University Press, Auckland, 1996.

C.A.B. Smith. The development of human linkage analysis. *Annals of Human Genetics*, 50:293 – 311, 1986.

M. W. Smith, J. A. Lautenberger, H. D. Shin, J-P. Chretien, S. Shrestha, D. A. Gilbert, and S. J. O'Brien. Markers for mapping by admixture linkage disequilibrium in African American and Hispanic popuations. *American Journal of Human Genetics*, 69:1080 – 1094, 2001.

R.S. Spielman, R.E. McGinnis, and W.J. Ewens. The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52:506 – 516, 1993.

B. K. Suarez, J. Rice, and Reich T. The generalized sib pair ibd distributions: Its use in the detection of linkage. *Annals of Human Genetics*, 42:87 – 94, 1978.

B.K. Suarez and C.L. Hampe. Linkage and association. *American Journal of Human Genetics*, 54:554 – 559, 1994.

J. Terwilliger and J. Ott. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *American Journal of Human Genetics*, 42:337–346, 1992.

J.D. Terwilliger and K.M. Weiss. Linkage disequilibrium mapping and complex disease: fantasy or reality? *Current Opinion in Biotechnology*, 9(6):578 – 594, 1998.

The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:925 – 933, 2001.

D. C. Thomas and J. S. Witte. Population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Prev Biomarkers*, in press, 2001.

R. Thompson. The estimation of heritability with unbalanced data. i. observations on parents and offspring. *Biometrics*, 33:485 – 495, 1977a.

R. Thompson. The estimation of heritability with unbalanced data. ii. data available on more than two generations. *Biometrics*, 33:497 – 504, 1977b.

P. J. Tienari, J. D. Terwilliger, J. Ott, J. Palo, and L. Peltonen. Two-locus linkage analysis in Multiple Sclerosis (MS). *Genomics*, 19:320 – 325, 1994.

H.K. Tiwari and R.C. Elston. Deriving components of genetic variance for multilocus models. *Genetic Epidemiology*, 14(6):1131 – 1136, 1997a.

H.K. Tiwari and R.C. Elston. Linkage of multilocus components of variance to polymorphic markers. *Annals of Human Genetics*, 61(3):253 –261, 1997b.

H.K. Tiwari and R.C. Elston. Restrictions on components of variance for epistatic models. *Theoretical Population Biology*, 54(2):161 – 174, 1998.

S. Van Gestel, J. J. Houwing-Duistermaat, R. Adolfsson, C. M. Cornelia M. van Duijn, and C. Van Broeckhoven. Power of Selective Genotyping in Genetic Association Analyses of Quantitative Traits. *Behavior Genetics*, 30(2):141 – 146, 2000.

T. van Wezel, A. P. M. Stassen, C. J. A. Moen, A. A. M. Hart, M. A. van der Valk, and P. Demant. Gene interaction and single gene effects in colon tumour suspeptibility in mice. *Nature Genetics*, 14:468 – 470, 1996.

C. Venter, M. D. Adams, E. W. Myers, and et al. The sequence of the human genome. *Science*, 291:1304 – 1351, 2001.

V. J. Vieland, D. A. Greenberg, and S. E. Hodge. Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. *Human Heredity*, 43:329 – 336, 1993.

J. Wang, R. Guerra, and J. Cohen. A statistically robust variance components approach for quantitative trait linkage analysis. *Annals of Human Genetics*, 62:349 – 359, 1998.

K. Weber, R. Eisman, L. Morey, A. Patty, J. Sparks, M. Tausek, and Z-B. Zeng. An analysis of polygenes affecting wing shape on chromosome 3 in Drosophila melanogaster. *Genetics*, 153:773 – 786, 1999.

K.M. Weiss and J.D. Terwilliger. How many diseases does it take to map a gene with snps? *Nature Genetics*, 26(2):151 – 157, 2000.

R. C. Williams, J. C. Long, R. L. Hanson, M. L. Sievers, and W. C. Knowler. Individual estimates of European genetic admixture associated with lower body-mass index, plasma glucose, and prevalance of type 2 diabetes in Pima Indians. *American Journal of Human Genetics*, 66:527 – 538, 2000.

J. F. Wilson, M. E. Weale, A. C. Smith, F. Gratix, B. Fletcher, M. G. Thomas, N. Bradman, and D. B. Goldstein. Population genetic structure of variable drug response. *Nature Genetics*, 29:265 – 269, 2001.

S. Wright. The genetical structure of populations. *Annal of Eugenics*, 15:323–354, 1951.

K. Yaffe, M. Haan, A. Byers, C. Tangen, and L. Kuller. Estrogen use, APOE, and cognitive decline: evidence of gene-environment interaction. *Neurology*, 54:1949–1954, 2000.